# 社交网络信息传播

张　熙　编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

## 内 容 简 介

社交网络信息传播是计算机科学、传播学、社会学、管理学等领域的重要研究问题，在舆情分析和网络营销领域具有广泛的应用。目前，同类著作更多地站在传播学或管理学角度介绍信息传播的模型、原理和应用。而本书主要从计算机科学角度出发，介绍了该领域的经典问题和最新成果，包括传播模型、话题检测、影响力最大化等问题。此外，本书面向实际应用场景，阐述了如何开发舆情分析和网络营销系统。

本书可供社交网络分析与数据挖掘研究领域的研究者了解该方向的前沿基础工作，也可供信息传播与网络舆情领域的工程实践人员作为系统构建的参考和指导。



**图书在版编目（CIP）数据**

社交网络信息传播 / 张熙编著. —北京：电子工业出版社，2016.8

ISBN 978-7-121-29783-0

Ⅰ. ①社… Ⅱ. ①张… Ⅲ. ①互联网络－信息－传播－研究 Ⅳ. ①G206

中国版本图书馆 CIP 数据核字（2016）第 205185 号

责任编辑：徐蔷薇
特约编辑：赵海军　赵海红等
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　　印张：13.75　　字数：220 千字
版　　次：2016 年 8 月第 1 版
印　　次：2016 年 8 月第 1 次印刷
定　　价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：xuye@phei.com.cn。

# 前　言

随着互联网进入 Web 2.0 时代，以新浪微博、网络社区、Twitter 和 Facebook 为代表的社交网络得到飞速发展，信息的传播速度更快、影响范围更广，正在深刻改变着人们的思维方式、行为模式和社会形态。深入理解社交网络中的信息传播模式和规律具有重要的科学价值，如能将其合理利用，将带来巨大的经济价值和社会价值。

社交网络信息传播涉及计算机科学、传播学、社会学、管理学和心理学等多个学科领域。目前，同类著作更多地站在传播学或管理学角度介绍信息传播的模型、原理和应用，而本书主要从计算机科学角度出发，基于近些年在数据挖掘和社交网络分析领域的研究经历与相关成果，系统梳理了社交网络信息传播的经典问题和最新研究成果。另外，面向实际应用中的需求，介绍了如何实现对传播信息和网络舆情的监测、分析和处理。

本书分为上、下两篇共 7 章。上篇从理论研究出发，第 1 章传播模型，介绍了社交网络中信息的两种传播模型，分别解释模型和预测模型；第 2 章热门话题检测，介绍了几种话题检测的算法，并结合实例进行了分析和对比；第 3 章影响力最大化，总结分析了几种社交网络影响力最大化传播模型及其优化算法；第 4 章收益最大化，介绍并分析了营销模型及策略，描述了相关的算法。下篇从工程实践出发，介绍了作者团队近年来开发的网络舆情监测系统（第 5 章）、品牌推荐和保护系统（第 6 章），以及其中涉及的一项核心技术——网站验证码识别（第 7 章）。

本书可供社交网络分析与数据挖掘研究领域的研究者了解该方向的前沿基础工作，也可供信息传播与网络舆情领域的工程实践人员作为系统构建的参考和指导。

感谢参与本书内容讨论、资料收集、内容编纂、成果贡献和审查校对的北京邮电大学可信分布式计算与服务教育部重点实验室的老师和同学：吴旭老师和颉夏青老师，博士生苏援和许晋，硕士生侯玉锋、李金兰和曲思宇，以及北京邮电大学国际学院的高炘、麦艺琼、吕浩然和郭鲲鹏同学。感谢“973”项目“社交网络分析与网络信息传播的基础研究”对本书的支持。

由于作者水平有限，书中难免有错误和疏漏之处，恳请读者批评指正。

# 目　录

## 上篇　理论研究

# 上篇

## 理论研究

# 第 1 章　传播模型

## 1.1　引言

近年来，随着社交网络的发展，对于信息传播模型的研究一直很活跃。本章介绍一些基本的传播模型，并描述这些模型如何推断出底层传播级联机制或预测消息传播过程。

在流行病学领域，对复杂系统中传染病传播过程的研究已经持续了几个世纪，例如，在某些条件下病毒增殖传播的预测。在线社交网络信息传播领域的研究也广泛地借鉴了流行病学的研究方法，但是过程更加复杂。我们不能直接套用传染病模型，一方面原因是在线社交网络规模非常大，网络更新和传播速度也更快，而很多原有模型的效率太低，难以实际应用；另一方面是用户类型和消息类

型更加多样，各个网络平台的传播规则也不尽相同，需要设计新的模型。

信息传播的预测具有广阔的应用场景，如市场营销、安全监测和网络搜索等。例如，对于市场营销来说，如果我们知道哪些特征主导传播过程，就可以更好地宣传产品或者保护其不受到网络攻击。同时，市场营销也可以通过合理选择初始投放广告节点来使得收益最大化，或者通过确定营销行动之间的时延来获利。另外，在安全维护的场景下，刑事调查员通常需要了解特定成员之间的信息流，以提取关于一个人或一群人是否有犯罪嫌疑的线索[16]。最后，对于 Web 搜索，一个传播预测模型可以帮助用户根据某话题热度的预期来增长订阅最热门的话题。这些都反映了传播预测模型的广泛作用。

传播过程的特征由两方面描述：第一个方面为结构，用一幅网络图描述出哪些节点间可以相互影响，网络的拓扑结构是怎样的；第二个方面为动态变化，如传播速率的演变，即在一段时间内接收某条消息的节点数量。

描述信息传播过程的基本方法是考虑网络中的一个节点是否可以被信息激活。因此，传播过程可以看作节点连续激活的序列，具体查看定义 1.1。

定义 1.1：（激活序列）网络中的一组有序节点连续接收某条消息，这组节点序列被称为激活序列。

在通常情况下，在线社交网络（Online Social Network，OSN）背景下的模型都只假设用户只接受相互连接邻居的影响。也就是说，一个 OSN 是一个封闭的世界，并且假设信息级联导致了信息的传播。这也是为什么网络中一条消息的路径通常被称为传播级联，如定义 1.2 所示。

定义 1.2：（传播级联）有向树的根可以作为激活序列的第一个节点，这棵树表示了节点之间的影响关系（有向边显示了信息传播方向），并且以激活序列依次展开。

激活序列如图 1-1 所示，黑色节点表示参与某一话题传播的活跃节点。但我们并不清楚这个消息如何传播及为何传播。因此，有必要建立模型来描述传播过程的底层机理。传播模型可以分为两类：解释模型和预测模型。

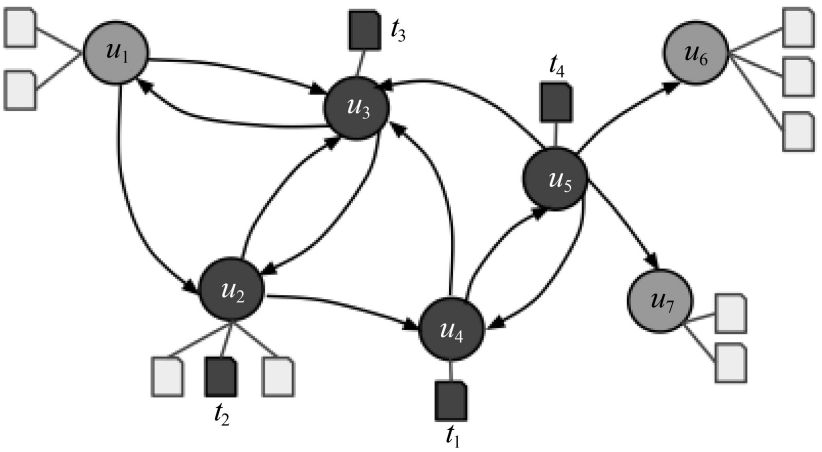



图 1-1　激活序列

本章主要内容安排如下：第二节介绍解释模型及相应算法；第三节介绍预测模型及代表性算法；第四节对两种模型进行总结。

## 1.2　解释模型

解释模型的目的是在给出完整的激活序列后，推断出底层传播级联机制。这些模型能够帮助我们了解消息是如何传播的。

### 1.2.1　问题描述

网络中的消息传播可以类比传染病扩散的过程。节点何时被感染往往可以直

接观察到，但是被谁感染却难以察觉。例如，我们可以知道某个人感染了感冒，但我们却很难知道是谁传染给他的。此外，在许多应用中，基础网络的传播扩散过程也是无法观察到的。为了应对这些挑战，需要开发出一种算法，追踪网络中的传播和影响路径，并且推断出整个传播网络；给出节点被感染的时间，找出能够解释所观察到的感染时间的最佳网络。由于推断网络的问题为典型的 NP-hard 问题，因此需要开发一种高效近似算法，扩展到大型实际数据集中，并且能够达到最佳效果。

为了研究网络信息传播，必须解决两个挑战。首先，为了跟踪网络中发生的级联过程，我们需要识别出在网络中通过边扩散和传播的信息，随后确定能够成功跟踪这些信息的方法。其次，传播发生的网络通常是未知的，通常只能观察到在哪些时间哪些节点受到感染，很难发现谁感染它们。

这些挑战在大型网络中尤为突出，而且过程复杂，影响因素多，缺少相关的系统性研究[2-5]。为了研究网络中的传播路径，需要了解完整的影响关系，如果在一个环节出错，则可能导致整个推断的错误。即使收集到了大规模的传播数据，在无须人工监督的情况下识别在传播过程中相对完整的文本片段依然是一项艰巨的任务。

### 1.2.2　解决方案

为了简单起见，我们假定基础网络是静态的，之后观察节点被感染（通过某些消息、商品或病毒）的时间（但不知道具体感染源）。因此，对于每个级联，首先观察节点被感染的时间，然后寻找消息在看不到的网络上的传播路径，最终目标是通过级联传播重组网络。

对于网络中的边有不同的解释：在病毒或灾难传播中可以解释为可能的感染路径，两个节点通过边相连代表病毒有可能通过边从一个节点感染另一个节点；

在信息传播中可以解释为复制；在病毒式营销中，边的概念为一个节点影响另一个节点。

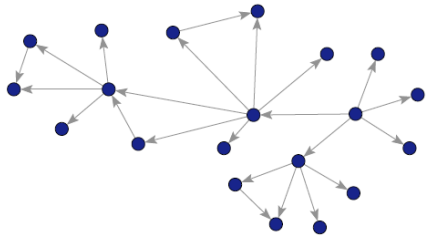
想要解决问题，前提是要观察不同节点间的传播级联，之后推断出网络中潜在的边。对于节点 $v$，如果在很多级联中 $v$ 在 $u$ 被感染之后而受到感染，那么可以推断在网络中$(u,v)$为一条边。通过探索节点受到感染的时间，最后目标是恢复无法观察到的传播网络。

### 1．NETINF 算法

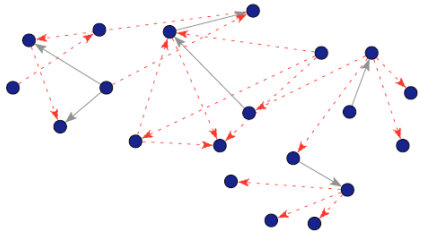
Gomez 等人提出了依据节点感染次序的相关性[1]，从而推断出传播级联的结构。假设活跃节点以一定概率影响其邻居节点，并且节点之间的影响相互独立。因此，一个节点已经向另一个节点传送消息的概率会随着距离激活时间变久而下降。基于上述假设，提出了可扩展的 NETINF 算法，它是一种基于子模函数优化的迭代算法，能够推断出传播网络。

具体过程是：首先构建一个概率模型，该模型需要解决在一个固定的假想网络中，级联是如何作为有向树来传播的（由于一个网络中的节点不会被重复感染，所以消息传播的路径中不存在回路，此时的消息传播就可以被视为一种树的传播形式）。由于只能观察到节点被感染的时间，可能存在多棵传播树来解释同一组数据，并且需要考虑所有的可能性。由于有大量不同情况的组合，该模型的计算需要指数时间复杂度，这样推断整个网络的传播级联的效率就很低，此问题为典型的 NP-hard 问题。因此，为了提高算法的效率，引入了近似的概念，有效优化目标函数。通过收益递减属性，可以证明 NETINF 算法接近最优网络。通过懒惰评估[6]，利用目标函数的局部结构能够加速 NETINF 算法。最终使得 NETINF 算法能够找到使得观测数据可能性最大化的传播级联。

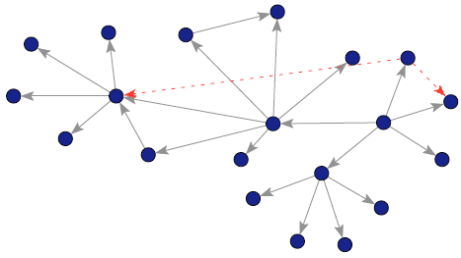
该算法下的传播模型与启发式算法下的传播模型对比如图 1-2 所示。



（a）真实传播模型



（b）启发式算法下的传播模型



（c）NETINF 算法下的传播模型

图 1-2　传播模型对比

### 2．NETRAT 算法

Gomez 等人[6]之后扩展了 NETINF 算法，称为 NETRAT 算法。该算法主要通过观察传染来推断底层传播过程。该方法首先需要构建一个模型，这个模型结合了生成传播扩散过程中基本的时空结构。生成模型需要做以下几点假设：第一，传播过程发生在静态（固定的）但未知的网络（有向图）中；第二，感染与否只有两种状态（1 或 0，二值型），即一个节点受感染或未受感染，没有部分感染概念或信息的部分传播概念；第三，感染沿网络的边彼此独立地发生；第四，感染可以发生在不同的时间，节点 $a$ 在时间 $t$ 感染节点 $b$ 的可能性取决于 $a$、$b$ 和 $t$ 之间的概率密度函数；第五，我们观察所记录的时间窗口中出现的所有感染过程。最终目的是推断网络的连通性，以及观察节点被感染后，推断通过边传播感染的可能性。

该概率传播模型的主要目的是描述在静态网络中一个传染过程在现实中是如

何发生的。通过观察感染级联来寻找最佳网络的问题可以简化为求解凸问题。凸问题可以分解成许多较小的问题，能够自然并行运算，使算法扩展到具有成千上万个节点的网络中。这种方法的主要创新之处在于，它将传播模型作为一个空间上的离散网络。感染传播底层机制具有相当的复杂性，例如，病毒对于人的感染性取决于天气、饮食、年龄、压力水平、之前是否接触过类似的病原体等。为了避免在建模过程中考虑背后的复杂因素，需要设计一个适用于大规模数据驱动的方法，该方法只能够根据可见的时空级联来推断传播过程。因此，只利用节点间的时间相关性、传输速率和感染时间来构建模型，而不依赖于未知的外部因素。此外，在以往的工作中，尚未对传播网络的持续时间进行动态建模，而这是理解传播过程的一个关键点。

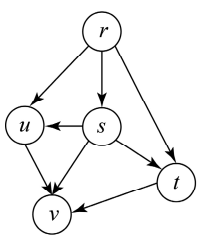
### 3．INFOPATH 算法

上面的算法假设基础网络不发生改变，但是这种假设并不符合实际，因为社交网络的拓扑结构演变得非常迅速，每时每刻都有边在创建或者删除。例如，一个博客上的一篇帖子被广泛传播之后，网络中会创建很多新的边，此后，这个博客上发布的内容也会有更强的扩散能力。类似地，在任何给定的时间内，可能会出现一个突发的事件或一个突发的话题，并在有限的一段时间内变得非常流行，这将再次导致新的边出现，或原有的边消失，这样会导致一个随时间变化的基础网络的诞生。我们通常用传染病的传播来类比信息传播的过程，在传染病传染的过程中，有因传染病而死亡的人导致减少了边和节点，还有因人口的出生导致增加了网络中的节点，这些都会导致网络拓扑为动态的、时变的。为了更好地理解这些时间变化，需要重构这些随时间变化的网络结构和基础时间动态，然后研究真实世界中事件、话题或内容的网络路径。

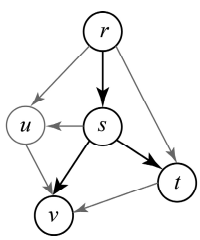
假设在观察网络中节点感染的时间时，还有很多观察不到的网络动态变化，如基础网络的边和边的动态变化。为了解决这个问题，Gomez 等人扩展了 NETRATE 算法，提出一个随时间变化的推论算法——INFOPATH，这种算法使用随机梯度来提供网络结构的在线估计和随时间变化的状态。

将传播过程建模为离散网络[1,6]，此模型仅记录节点感染时间和传播的事件，允许信息通过数据驱动的方法以不同的速率在网络中不同的边缘传播。该模型忽略了外部消息来源，而仅仅考虑网络中内部消息的传播过程[7]，采用时变网络推理算法——INFOPATH 能够提供在线社交网络空间上的时间演化。

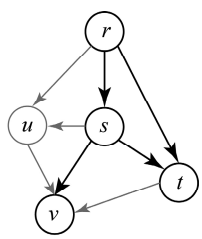
**k-tree**：在很多情况下，因为难以看到节点上执行的所有操作，所以级联模型可能是不完整的，也就是说，存在数据缺失。缺失数据对于网络的推测有着严重的影响，比如在潜在受种疫苗对象的检测或者在博客圈中预测信息流向时，如果缺失的是网络中的核心数据，就会对上述应用的结果产生很严重的影响。如果把网络图形化，在图 1-3 中，（a）显示了一张网络图；（b）描述了一个影响力级联；（c）描述了一个网络级联，只能从中观察到参与活动的节点，而不能观察到传播的边缘，边缘可以基于时间顺序从网络上推断；（d）和（e）分别显示了缺失数据的影响力级联和网络级联。缺失数据的级联可能不再具有与原始级联相同的属性（如深度、边的数量），甚至可能不再连通。那么，需要通过观察部分级联来估计完整级联。



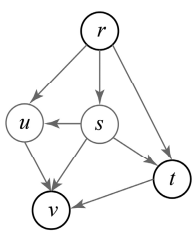

（a）网络模型



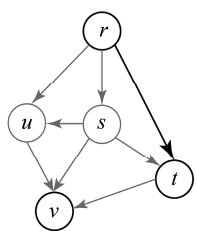

（b）影响力级联




（c）网络级联




（d）缺失数据的影响力级联




（e）缺失数据的网络级联

图 1-3　级联中缺失的数据

级联缺失有很多原因。大多数社交网络不提供用户活动的全部信息，因此只能观察参与级联用户的一个子集。此外，基于隐私考虑，很多用户不愿意共享他们的数据。另一个原因是收集完整信息的费用太高。总之，随着社交网络本身的快速增长，生成的数据量日益增加，以及对用户隐私的保护升级，数据丢失的问题越来越显著。

在缺失数据的信息级联方面的研究较少。Choudhury 等人[8]考虑了不同取样策略对传播的测量性能的影响。Sadikov 等人[9]开发了一种基于 k-tree 模型的方法，通过完整级联序列的一小部分，估计完整级联的属性，如它的大小或深度。这种方法需要考虑均匀随机抽样，每个节点的丢失概率相互独立。这种方法不仅能够解释通过采样产生的失真，而且能够校正失真（例如，推断出完整的级联性质）。校正失真对于级联树状图具有很大的挑战，因为其在很多情况下非常脆弱，容易与缺少节点的一小部分结构断开。

## 1.3 预测模型

预测模型旨在通过从过去时间、空间上的传播轨迹来学习如何在一个给定的网络中预测一个特定的传播过程将如何展开。现有的模式可以分为两大方向：基于图形的方法和基于非图形的方法。

### 1.3.1 基于图形的方法

基于图形方法的模型有两种：独立级联模型（IC）和线性阈值模型（LT）。两种模型都假设传播扩散过程是在一张静态结构图上进行的，并且都是基于有向图的，其中每个节点根据模型条件设定来判断是否被激活，并且激活后的节点会保持激活状态。不同之处在于，IC 模型需要知道每条边的传播概率，而 LT 模型

需要对每条边定义一个影响程度并对每个节点定义一个影响阈值。这两种模型都从一组初始节点启动，沿着时间轴以同样的方式重复执行。

在 IC 模型下，对于每次迭代，新激活的节点都会有一次机会以一个特定概率去激活其邻居节点，特定概率为两个节点之间的边上所定义的概率；在 LT 模型下，对于每次迭代，不活跃节点通过活跃邻居节点对其总影响程度来决定是否被激活，若总影响程度超过节点阈值，则此节点被激活。节点被激活后，在下一次迭代时会变为活跃状态。在这两种模型下，当没有新的节点被激活时，传播过程结束。这两种机制反映了两种不同的观点：IC 是以发送者为中心的，而 LT 是以接收者为中心的。图 1-4 展示了 IC 传播模型的一个例子，后面将详细介绍这些模型及它们在社交网络中的应用。

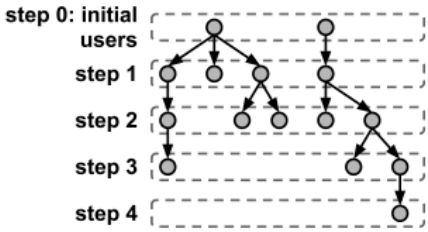



图 1-4　独立级联模型下的 4 步传播过程

关注与被关注的社交网络是一种独特的、具有活力的通信媒介，并且越来越多地成为病毒营销的宿主，如突发新闻的传播、紧急广播、营销、公关、宣传行动等。典型的是微博（或 Twitter），对其合理建模并理解其中的信息流才能够更有效地对其加以利用。在微博（或 Twitter）上一个常用的信息共享方式是网址（URL）。当用户在推文中发出一个有趣的 URL 时，关注此用户的人有可能转发这个消息来让更多的人知道[11]。

LT[10]模型侧重于对 URL 转发形成消息级联进行特征观察和建模。在以前的工作中，建立的大多数信息传播模型要么重现与经验观察相匹配的统计特性级联，要么在给出一些根节点后预测信息在网络中会传播多远。而此模型解决了另一种

问题：根据现有 URL 转发的训练集，预测哪些用户将转发哪些 URL。

URL 转发的精确预测是很多应用的重要推动者。首先，当知道每个用户和 URL 发出推文的概率后，可以对每个用户生成一张 URL 列表，这张列表能够根据用户的兴趣对用户提供个性化推荐。对于关注了很多其他用户的用户，此法可用于防止信息过载并过滤一些不重要的消息。其次，汇总每个 URL 的概率来量化 URL 在社交网络中的传播潜力，可以作为对病毒链接的早期鉴定。再次，一个特定社交网络训练的精确的传播模型能够帮助病毒营销策略选择起始 URL 并使得最终传播范围最广。最后，预测模型不仅当它的预测正确时是有用的，而且在与新数据不匹配时也是有价值的，因为网络中突然爆发的一些异常活动往往预示着特殊值得关注的事件，例如垃圾邮件检测。

LT 传播模型同时考虑到了几个关键因素：内容流行程度、用户影响力和传播速度，这几个因素都是此模型的预设参数。在此模型中使用梯度上升的方法找到使得数据预测正确率最大的参数值，结果验证该模型能够预测出一半以上的 URL 转发，只有 15%的失误率，效果如图 1-5 所示。

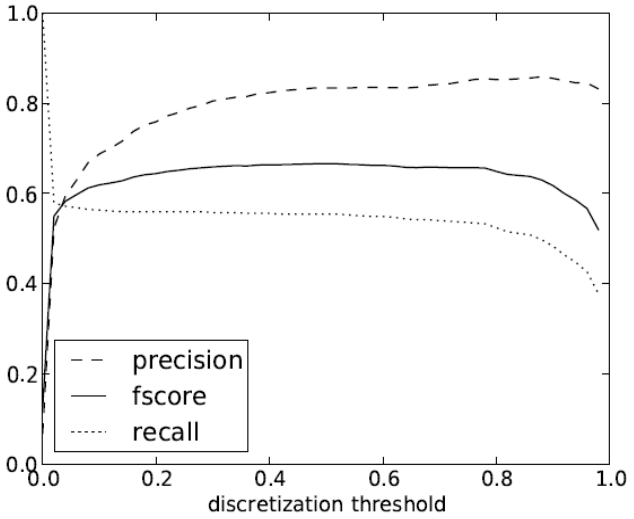



图 1-5　LT 模型下的 precision、recall 和 fscore

IC 模型与 LT 模型关注的是信息传播的不同层面。IC 模型是以发送者（受感染者）为中心的，描述的是一个发送者（受感染者）主动向它的邻居节点“推”（发出）消息，每个活跃节点都以一定概率激活周围的非活跃节点。而 LT 模型是以接收者为中心的，节点监听网络信息，从邻居节点中“拉”（接收）消息，每个节点都有一个接收阈值，当连通该节点的级联上的总权重超过该阈值时，该节点就被激活。

IC 模型和 LT 模型都需要一些预设值，如 IC 模型的节点的信息传播概率、LT 模型的边的权重及每个点的阈值。这些值需要预先知道，但是在实际中，这些变量是很难检测与量化的，于是就需要通过一些训练数据及机器学习的方法来将这些参数计算出来，从而导致预测模型的准确度受到机器学习算法及训练数据优劣的影响。

**AsIC 模型[12]：**IC 模型和 LT 模型还存在一个问题，它们把信息传播看作节点的一系列状态变化，并且这些变化都是以同步方式进行的，这种方式等同于假设一个离散的时间步骤。然而，实际的传播却沿着连续时间轴以异步方式进行，并且观察到的数据的时间标记并非等距。因此，有必要扩展这两个模型使其能够模拟异步时间延迟（扩展模型为 AsIC 模型和 AsLT 模型）[13-15]。还有其他工作也试图通过类似思想推断底层网络[3,8]。采用 AsIC 模型也能解决同样的问题，但与上述研究不同的是，此模型学习节点上传播概率的相互依赖性和时延参数的属性，而不是直接从所观察到的数据中学习。

AsIC 模型如下：令 $G=(V,E)$为一个没有自链接的有向图，$V$ 代表节点集合，$E$ 代表边集合。对于每个节点，令 $F(v)$表示 $v$ 的所有出边相连的节点，$B(v)$代表 $v$ 的所有入边相连的节点。节点接收某一消息表示被此消息激活，激活为单向的，不会反向发生。

AsIC 模型有两个参数：$S_{\max}$ 和 $r_{u,v}>1$，对于每条边$(u,v)\in E$，$P_{u,v}$ 为传播概率，$r_{u,v}$ 为链路的时延参数。在连续时间 $t$ 内展开信息扩散过程并且在给定初始活动节

点后展开下列过程：当一个节点 $u$ 在 $t$ 时刻被激活时，它将会有一个机会在 $t+\delta$ 时刻之前激活每个非活跃节点 $v$，$\delta$ 为从 $r_{u,v}$ 分布中随机选择的时延。如果 $u$ 成功，则 $v$ 在 $t+\delta$ 时刻变为活跃状态；如果没有节点可以被激活，则过程结束。IC 模型和 AsIC 模型对比如图 1-6 所示。

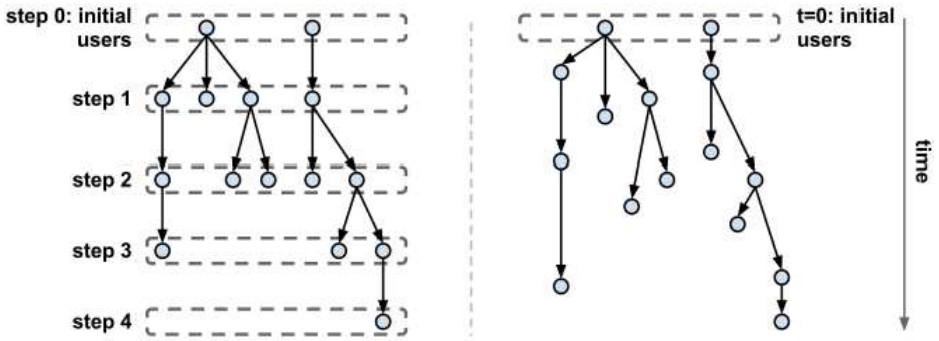



图 1-6　IC 模型和 AsIC 模型

在现实中，扩散概率和所述网络中的链路时延参数必须至少是两个相互连接的节点的属性函数，忽略此属性则不能反映实际情况。使用这种关系的另一大优势是能够避免过拟合的问题。由于边的数量远远大于节点的数量，即使已知社交网络的稀疏性，需要学习的参数数目依然是巨大的，需要通过海量的数据来分别学习每个扩散概率。为了提高效率，许多研究都假定在不同的链接中该参数是一致的，或者它仅依赖于话题。然而相对于学习每个扩散概率，学习一个函数更加现实，而且不需要如此庞大的数据量，参数更新算法也被验证保证收敛。该算法能够正确估算出传播概率和时延参数，相对于一致同步时延的假设，更加符合实际。

**T-BaSIC 模型：**一个消息在实体间传递，即在封闭环境中的社交网络用户在忽略外部影响的情况下有可能会接受这条消息。一个传播扩散过程可以和一个拓扑结构相关联[17]，如规模、范围和事件属性等。此模型要解决的问题是：在一个封闭环境中，用户通过社交网络交互，如何模拟这种环境并且预测传播特性。

T-BaSIC 模型主要考虑传播扩散过程的时间动态，能够从一种更实际的角度出发通过机器学习的算法建立模型，预测社交网络的信息传播过程。它假设社交网络中的信息传播依赖于用户之间的连接图，并且是根据局部特性由节点之间的微相互作用解释的。之后根据图中个人的行为进行统计分析，而非全局行为。

### 1.3.2　基于非图形的方法

基于非图形的方法不需要一个特定的消息传播图形结构的存在，此方法以往主要应用于对流行病学过程建模。非图形方法将节点分成几类（状态），并将关注点放在每一类节点比例的变化上。SIR 和 SIS 是两个开创性的模型[18,19]，其中 S 代表“敏感（或易感染）”，I 代表“感染”（接受了某信息），R 代表“恢复”（从感染中恢复）。在这两个模型下，在 S 状态的节点会以固定概率变为状态 I。然后，在 SIS 情况下，在 I 状态的节点会以固定概率变回状态 S；而在 SIR 情况下，它们将永久变为 R 类，表示被治愈，并已具有免疫力，不会再被感染。这两个模型假设每个节点连接到另一个节点的概率都相同。

**SIS 模型[20]：**需要一个生成模型来生成级联，SIS 的目标是找到最简单、直观的模型，并且参数尽可能少。Leskovec 等提出了一个简单而直观的 SIS 模型，仅需要一个参数。它假定所有节点都以相同的概率 $\beta$ 被感染（激活），被感染的节点在下一时间段重新成为敏感节点。

以博客为例说明级联的产生过程。在某个博客上发布了一篇帖子，其他博主阅读后会发表一些其他文章，并将原链接附于其后，此过程继续进行，并形成一个级联。博客的级联类似于病毒（信息）在网络中的传播，即初始帖子负责感染之后的博客。由于级联过程的存在，病毒（信息）在网络上传播，并留下痕迹。为了模拟这一过程，用一个参数 $\beta$ 衡量博客上帖子的感染力。

下面描述 SIS 模型。每个博客都有两种状态：被感染或敏感。如果博客处于受感染的状态，这意味着博主刚发布了一篇帖子，此博客现在有机会扩大其影响

力。只有在敏感（未被感染）状态的博客可以被感染。当一个博客成功感染了另一个博客后，则在级联中添加一个新节点，新感染节点和感染源之间产生一条边。之后该节点马上恢复，即一个节点只在一个时间步长内保持受感染的状态。这样能够使得模型有能力多次感染其他博客，也就是一个博客中的多篇帖子能参与到同一级联中。

下面是对该模型假设的一些分析。首先，可以看到博客受感染后会立即恢复，因此可以被感染多次。每次有博客被感染就会有新节点加入到级联中，说明一个博客中的多篇帖子可以加入同一级联。其次，可以看到在这个模型中，没有区分特定话题或特定博客的影响力，所有博客的参数都是一样的。最后，上述过程所产生的级联是树形的。现实世界中大部分的级联都是树或类似树的形式。同一参数是一个很强的假设，因为在现实世界的社交网络中，并非所有节点之间都具有相同的影响，例如有些大 V 的影响力比普通用户的影响力更强，不同内容的信息的传播也会对传播概率等参数产生影响，因而导致节点之间有不同的影响参数，于是就有必要设计更复杂的模型。

**LIM 模型[21]：**当使用上述模型来描述现实世界中的网络传播时，通常需要做出以下几个假设：（1）我们可以收集到完整的网络数据；（2）消息只能在底层网络的边上传播；（3）网络结构足以解释观测到的行为。然而，在许多情况下，网络上发生的传播行为实际上是隐式的，甚至是未知的。下面对现实世界中的网络进行建模。

此模型假设整个信息传播的过程是由单个节点的影响力支配的。该方法主要通过对信息扩散的时间动力学进行预测，描述一条信息的传播扩散速率，通过建立线性影响模型（LIM），让单个节点的影响函数控制整个扩散速度。图 1-7 表示 LIM 如何从一组初始节点和它们的激活时间来预测传播扩散速度。LIM 通过把一组给定初始节点的影响函数相加来预测传播扩散速度。在这里，初始节点是 $u_1$、$u_2$ 和 $u_3$，其各自的影响力函数为 $I_{u_1}$、$I_{u_2}$ 和 $I_{u_3}$。

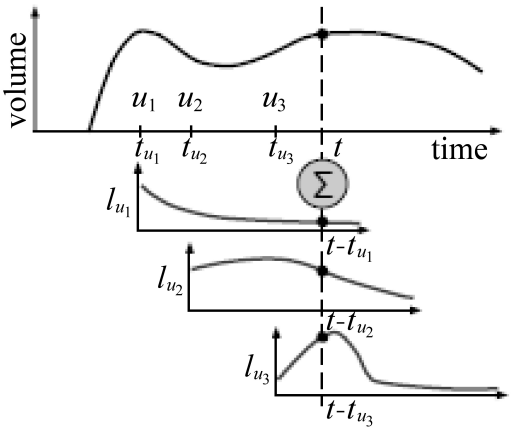



图 1-7　通过 LIM 预测传播扩散速度

LIM 模型关注于模拟一个节点对全局的影响力，而不是预测哪个节点会感染其他节点，或是节点感染哪一个节点。此扩散模型通常忽略时间和在离散时间段内的工作，但是准确模拟了每个节点的影响力并且描述了整个传播扩散过程。

LIM 模型假设新受到感染的节点数量取决于之前哪些节点受到感染，在这个模型中，每个节点又有一个与其相关的影响函数，在时刻 $t$ 新感染的节点数量是关于在 $t$ 之前受到感染的节点数量的影响函数。在信息传播中，转发某信息的节点数量取决于之前转发此信息的节点。此节点的影响力函数为：当节点 $u$ 在时刻 $t$ 转发某消息后，会导致在下一个时间段上其他 $I_u(1)$个额外节点转发此消息，在两个时间段后其他 $I_u(2)$个额外节点转发此消息。对于每个节点 $u$，得出一个 $I_u(t)$ 函数，在时刻 $t$ 所有节点会受到已被感染节点的影响，因此节点影响函数能够通过建立一个回归任务来完成。LIM 模型的影响函数是以非参数方式实现的，并且能够利用最小二乘法来进行估计。

**DL 模型[34]：**现有工作大都集中在对网络结构、用户相互作用和传播特征的研究上[21-25]。很多工作已经利用经验方法研究了在线社交网络的信息传播特点[26-30]，进而测量随时间推移的评论数量的变化，或某消息经过的平均链接数量。最近很多研究利用数学模型来预测社交网络中的信息传播[31,32]。然而，在时间和

空间上对传播模型深入理解的研究却较少。

社交网络中的信息传播过程如图 1-8 所示。

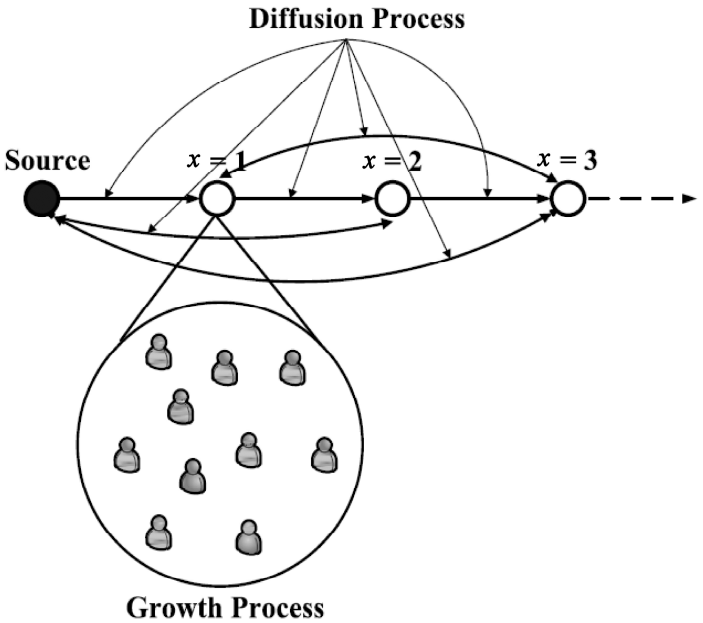



图 1-8　社交网络中的信息传播过程

利用 DL 模型来研究在线社交网络中信息传播过程的时间和空间模式，希望能够回答时空扩散问题：对于一个从特定用户 $s$ 发起的信息 $m$，在之后的时间段 $t$ 内，在距离 $x$ 处的地方受到影响的用户密度是多少（一个受影响的用户是喜欢该信息的用户）？解决时空扩散问题，有助于预测类似信息的传播模式。

DL 模型将在线社交网络的信息扩散过程分成两个独立的过程：成长过程和扩散过程。成长过程表示了从信息源开始，信息以相同距离在用户间传播的过程[33]；扩散过程表示了从信息源开始，信息以不同距离在用户间随意传播的过程[33]，扩散过程使用菲克扩散定律来衡量[9]。理论分析表明，此模型具有严格的递增性。此特性使 DL 模型成为最好的社交网络用户影响力密度模型。

## 1.4　本章小结

表 1-1 为解释模型的总结。

表 1-1　解释模型的总结

| 方　　法 | 网络特性 | | 推断性质 | | | 支持数据丢失 |
|---|---|---|---|---|---|---|
| | 静态 | 动态 | 成对传输概率 | 成对传输速率 | 级联概率 | |
| NETINF | √ | | √ | | √ | |
| NETRATE | √ | | √ | √ | √ | |
| INFOPATH | √ | √ | √ | √ | √ | |
| k-tree | √ | | | | √ | √ |

表 1-2 为预测模型的总结，主要识别了传播过程中的节点角色和有影响力的节点。

表 1-2　预测模型的总结

| 方　　法 | 维　　度 | | | 依　　据 | | 数学模型 | |
|---|---|---|---|---|---|---|---|
| | 社会 | 时间 | 内容 | 基于图形 | 基于非图形 | 参数模型 | 非参数模型 |
| LT-based | √ | | √ | √ | | √ | |
| AsIC | n/a | n/a | n/a | √ | | √ | |
| T-BaSIC | √ | √ | √ | √ | | √ | |
| SIS-based | | √ | | | √ | √ | |
| LIM | √ | √ | | | √ | | √ |
| DL | √ | √ | | | √ | √ | |

本章介绍了两种类型的社交网络传播模型：解释模型和预测模型。关于预测模型，介绍了基于图形和非图形的方法，基于图形的方法能够预测谁会影响谁，但不能在网络是未知的情况下使用；而基于非图形的方法忽略了网络的拓扑结构，只预测信息在全局范围内的扩散概率。

# 参考文献

[1] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of di↵usion and influence. In KDD ¡¯10, pages 1019¨C1028, 2010.

[2] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In Web Intelligence,2005: 207-214.

[3] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In SDM’07: Proc. of the SIAM Conference on Data Mining, 2007.

[4] J. Leskovec, A. Singh, and J. M. Kleinberg. Patterns of influence in a recommendation network. In PAKDD ’06: Proc. of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2006: 380-389.

[5] D. Liben-Nowell and J. Kleinberg. Tracing the flow of information on a global scale using Internet chain-letter data. Proc. of the National Academy of Sciences, 2008, 105(12):4633-4638.

[6] M. Gomez Rodriguez, D. Balduzzi, and B. Sch¨olkopf. Uncovering the temporal dynamics of di↵usion networks. In ICML ’11, 2011: 561-568.

[7] S. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In KDD ’12: Proc. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

[8] M. de Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In ICWSM ’10.

[9] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In WSDM ’11, 2011: 55-64.

[10] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In WOSN ’10, 2010: 3-11.

[11] DANAH BOYD, GOLDER, S., AND LOTAN, G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In Proceedings of HICSS-42, 2010.

[12] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda. Learning diffusion probability based on node attributes in social networks. In ISMIS ’11, 2011: 153-162.

[13] Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: ACML2009. 2009: 322-337.

[14] Saito, K., Kimura, M., Ohara, K., Motoda, H.: Behavioral analyses of information diffusion models by observed data of social network. In: SBP10. LNCS 6007, 2010: 149-158.

[15] aito, K., Kimura, M., Ohara, K., Motoda, H.: Selecting information diffusion models over social networks for behavioral analysis. In: ECML PKDD 2010. 2010: 180-195.

[16] A. Bennamane, H. Hacid, A. Ansiaux, and A. Cagnati. Visual analysis of implicit social networks for suspicious behavior detection. DASFAA’11, 2011.

[17] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. ICWSM’10, 2010.

[18] H. W. Hethcote. The mathematics of infectious diseases. SIAM REVIEW, 2000, 42(4):599-653.

[19] M. E. J. Newman. The structure and function of complex networks. SIAM Review, 2003, 45:167-256.

[20] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In SDM ’07, pages 551-556, (short paper) 2007.

[21] J. Yang and J. Leskovec. Modeling information di↵usion in implicit networks. In ICDM ’10, 2010: 599-608.

[22] T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. SIAM J. on Optimization, 1996, 6(4):1040-1058.

[23] F. Wang, H. Wang, and K. Xu. Diffusive logistic model towards predicting information diffusion in online social networks. In ICDCS ’12 Workshops, 2012: 133-139.

[24] A. Nazir, S. Raza, D. Gupta, et al..Network Level Footprints of Facebook Applications. in Proceedings of ACM SIGCOMM International Measurement Conference, November 2009.

[25] A. Nazir, S. Raza and C.-N. Chuah. Unveiling Facebook: A Measurement Study of Social Network Based Applications. in Proceedings of ACM SIGCOMM International Measurement Conference, October 2008.

[26] K. Lerman and R. Ghosh. Information Contagion: an Empirical Study of Spread of News on Digg and Twitter Social Networks. in Proceedings of International Conference on Weblogs and Social Media (ICWSM), May 2010.

[27] G. V. Steeg, R. Ghosh and K. Lerman. What Stops Social Epidemics?. in Proceedings of International AAAI Conference on Weblogs and Social Media, July 2011.

[28] S. Ye and F. Wu. Measuring Message Propagation and Social Influence on Twitter.com, 2010.

[29] M. Cha, A. Mislove, B. Adams, et al..Characterizing Social Cascades in Flickr. in Proceedings of ACM SIGCOMM Workshop on Social Networks (WOSN), August 2008.

[30] M. Cha, A. Mislove and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. in Proceedings of the 18th international conference on World wide web, ser. WWW ’09, 2009.

[31] J. Yang and J. Leskovec. Modeling Information Diffusion in Implicit Networks. in Proceedings of IEEE International Conference on Data Mining, December 2010.

[32] R. Ghosh and K. Lerman. A Framework for Quantitative Analysis of Cascades on Networks. in Proceedings of Web Search and Data Mining Conference (WSDM), February 2011.

[33] J. D. Murray, Mathematical Biology I. An Introduction. Springer-Verlag, 1989.

[34] F. Wang, H. Wang and K. Xu. Diffusive logistic model towards predicting information diffusion in online social networks. In ICDCS ’12 Workshops, 2012: 133-139.

# 第 2 章　热门话题检测

## 2.1　引言

在信息传播领域的研究中，热门话题检测是一项重要的任务。主要有两个目标：第一是对话题热度发展提供一个全局视图；第二是预测新的热门话题，主要包括提取“话题表”来总结话题、向用户推荐热门话题及预测未来热门话题。

然而传统的热门话题检测技术是用来分析静态语料库的，主要考虑一个固定时间段内话题的热门程度变化，而忽略了这段时间前后的行为，难以适应实时变化的在线社交网络消息流，而在线社交网络需要从消息流中检测突发话题。突发

话题是指在一段时间内引起了广泛关注，但是在这段时间之外并未得到大量关注的话题。

为了有效检测在线社交网络文本流中的话题，可以将重点放在突发话题上。检测突发话题需要依靠频率计算和离散数据，因此需要将消息流离散化，即将原始连续数据转换为同样大小的时间片段上的消息序列，如图 2-1 所示。

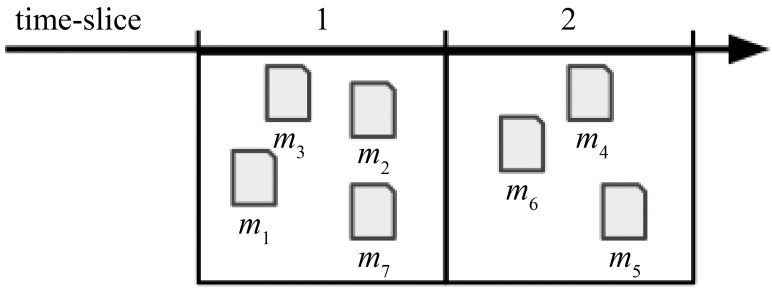



图 2-1　消息序列片段

这种离散过程可以在短时间片内检测热门话题而不可实现长时间片的检测。

本章主要内容安排如下：第一节对突发话题的概念进行简要介绍；第二节到第六节分别介绍话题检测的 PT 模型、OLDA 模型、TSTE 模型、SDNML 模型和 MACD 模型。下面将详细介绍从消息流中检测突发话题的方法。

## 2.2　热点话题（PT）模型

本节主要介绍热点话题（PT）模型[1]。随着时间的推移，微博中的话题也不断随之发生变化，为了在微博中找到根据时间而变化的热门话题，本节主要提出两种方法：第一，通过找到“热点话题”来提取有效的目录；第二，找到能持续很久但并非受到广泛关注的话题，即“持续性话题”。

### 2.2.1 热点话题简介

当微博涉及特定事件时，这些博文可以用于研究事件的结构。如在体育赛事和电视直播期间，人们会第一时间在网上做出反应。在这些微博流中，可以调查仅能一时引起人们兴趣的瞬时话题和在整个数据流中长时间存在的话题趋势。具体来说，有两个目标。

（1）热点话题：高度局部性，并且只引起人们一时兴趣的话题。

（2）持续性话题：能维持长时间的话题。

为了挖掘这两个指标，可以使用一种简单的话题评分方法，这种方法类似于TF-IDF 模型[2]。在 TF-IDF 模型中，一个特定文档中某话题的显著性被定义为该话题在文档中出现的次数（话题频率，TF），被该话题所在的文件总数（逆文档频率）归一化。传统上来讲，每篇博文是一个文件，并且 TF-IDF 可以在每篇博文的每个话题上返回一个唯一的分数。

### 2.2.2 热点话题

热点话题是指在某一特定时间段内非常热门的话题，即只在这段时间保持高热度而并不会延续到其他时间窗口的话题。首先在一个时间窗口内检查各种不同话题的频率并根据“时间窗口话题频率”（$tf_t,i$）为话题评分，之后通过“语料库话题频率”（在整个文件集合中包含话题 $i$ 的博文总数 $cf_i$）来归一化这个值。通过这两项计算，可以得出“标准化词频”为 $ntf_{t,i}=\frac{tf_{t,i}}{cf_i}$，此公式可以直观描述在时间窗口 $t$ 内包含话题 $i$ 的博文总数的百分比。为了便于试验，可以设置滑动时间窗口为 5 分钟（$t$ 点正负各 2.5 分钟）。

可以预见到，在某些能够引起大众强烈兴趣的时间点上，博文上此话题的频

率非常高，然而在其他时间频率会相对较低。为了自动找到这样的时刻，根据每个话题的尖峰度将其排名。尖峰度为话题 $i$ 的 $ntf_{t,i}$ 最大值。如果话题 $i$ 在 $t$ 时刻达到明显峰值，则可以推断在时间窗口 $t$ 内发生了一件令人感兴趣的事，而这个话题是对那件事的反映。

### 2.2.3　持续性话题

除了流行趋势下的突发话题，还需要找到持续性话题，即没有很强的突发性但能持续一段时间的话题。为了自动获取这样的时刻，首先需要找到时间点 $t_{\text{peak},i}$，这个时间点为每个话题 $i$ 的归一化话题频率峰值所在的时间点。对于一个能够持续引起大众兴趣的事件，这个话题在 $t_{\text{peak},i}$ 之前使用频率较低，而之后会被更频繁地使用。在评价这个方法时，先计算每个话题分别在 $t<t_{\text{peak},i}$（峰前）和 $t>t_{\text{peak},i}$（峰后）的 $ntf_{t,i}$ 均值；接着通过计算峰后均值与峰前均值的比值来评价其持久度；最后根据所有话题得分进行排序。

### 2.2.4　模型应用

下面用 Twitter 上的两个数据集来试验上述两个度量指标。第一个数据集包含了 53 712 个关于奥巴马就职典礼的博文，是一个十分具有代表性的样本。收集数据的时间为美国东部时间 2009 年 1 月 20 日 11:30—13:00。第二个数据集包含了 110 万个关于 MTV 音乐录影带大奖（VMAs）的博文，是一个具有完整性的样本。收集数据的时间为美国东部时间 2009 年 9 月 13 日 20:30 到 9 月 14 日 00:30（13 日晚 12:30）。

**应用一：奥巴马就职典礼**

奥巴马就职典礼的进行时间为 12:00—12:30，总统宣誓则在 12:05—12:07，在 12:25 左右就职演说结束。图 2-2 展示了具有最高归一化话题频率分数的词条，

每个词条都反映了就职典礼中的实际进程。“aretha”、“yoyo”和“warren”分别反映了 Aretha Franklin、Yo-Yo Ma 和 Rick Warren 的现身，“booing”表示 George W. Bush 的出现，当他乘坐直升机离开时产生了词条“chopper”，排名最高的词条“remaking”对应奥巴马的演说，“anthem”表示奏国歌。

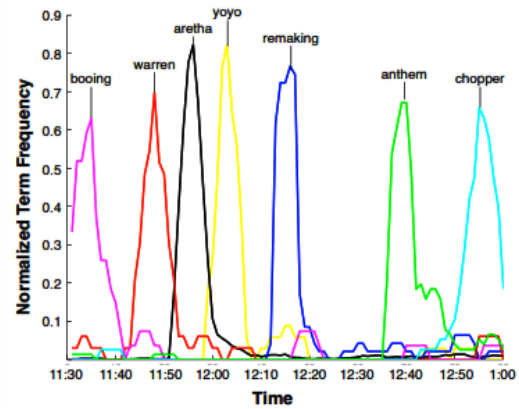



图 2-2　2009 年奥巴马就职典礼热点词条

图 2-3 显示人们持续关注时间最长的两个词条，分别是“flubbed”和“messed”，二者都与首席大法官 Roberts 说话时错误地交换了几个词的次序有关。然而，与图 2-2 中热点词条的峰值不同，它们在事件发生后相对长的一段时间内还被多次使用。

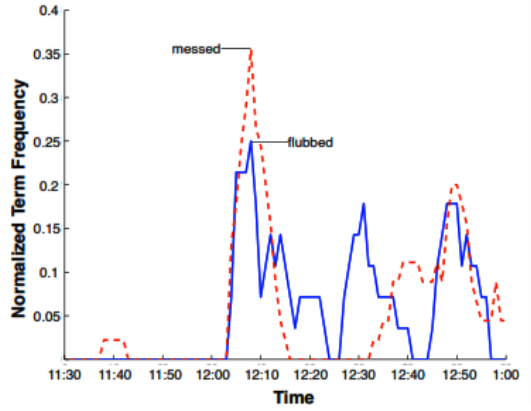



图 2-3　在 2009 年奥巴马就职典礼上被持续使用的两个词条

### 应用二：MTV 音乐录影带大奖

下面应用归一化热点词条频率分析 VMA 数据集，找出热点词条和人们的兴趣点，结果如图 2-4 所示。可以再次看到 Twitter 的话题趋势反映了事件的演变过程。在活动即将开始阶段，可以看到类似“firetruck”和“carriage”的词条，这些词条对应演员、歌手的车到达现场。在颁奖晚会上，可以看到反映登台的主持人、演出人员或者正在演唱的歌曲的趋势。如“thriller”在 Janet Jackson 向 Michael Jackson 致敬的表演中达到了峰值。“perrys”对应 Katy Perry 和 Joe Perry 共演。“furtado”的峰值出现在 Nelly Furtado 获奖时。在活动接近尾声，Taylor Swift 被 Kanye West 打断后重新发表获奖感言时，词条“noble”出现峰值。

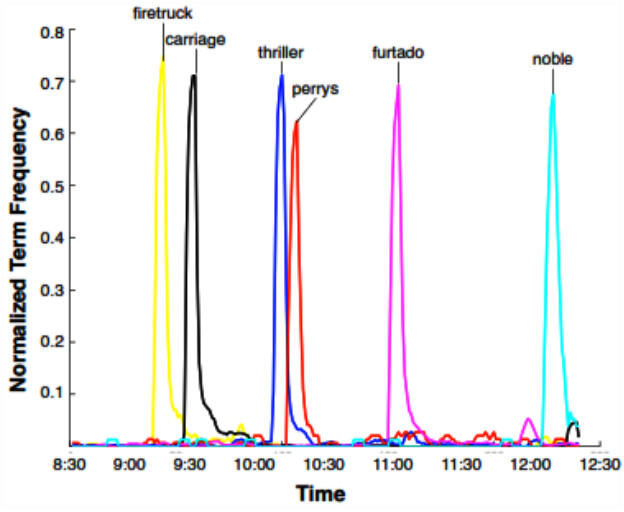



图 2-4　2009 年 MTV 音乐录影带大奖最具热度的词条

如图 2-5 所示，West 打断 Swift 的话题也是持续最久的话题。在 10:30，当 West 拿起麦克风后不久，“kanye”一词和 West 其他称呼的使用频率开始在 Twitter 上激增。人们对此事的持续兴趣反映在不间断的讨论及在此后数周中出现的各种各样的网络语言化身中。这种情况再次表明，在 Twitter 上长期讨论的话题会保持下去，并且热度超过了事件发生的时刻。

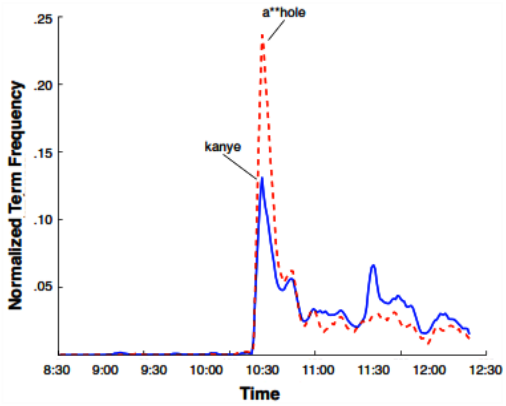



图 2-5　2009 年 MTV 音乐录影带大奖被持续使用时间最长的两个词条

## 2.3　在线话题模型（OLDA）

本节主要介绍在线话题模型（OLDA）[3]，此模型能够自动捕捉主题模式，并且能够识别文本流中出现的话题及其随时间的变化。这种方法可以使得话题模型框架，特别是 LDA（Latent Dirichlet Allocation）模型[4]实现在线工作，使其在有新文档（或文档集）生成时能够逐步建成一个即时更新的模型（每个文档话题的混合及每个话题词汇的混合）。基于经验贝叶斯的方法可以解决此问题，首先在无须访问先前数据的情况下，以从新数据流中推断出来的信息来更新当前模型。该方法具有动态性，并且便于实时跟踪随时间变化的主题并实时监测新出现的主题。

### 2.3.1　概率话题模型和 LDA 模型的应用

随着时间流电子文件的出现，时间信息变得必不可少，其内容包含了很强的时间序列。将时间信息纳入考虑范围可以更好地了解潜在话题并跟踪其随时间的演变和传播。此模型并不是分析线下收集到的时间戳文件集合，而是对当前到达

的文本数据流进行分析、总结和分类，这样的工作更加具有实用性。此外，更大的潜力在于依靠自动化系统来跟踪人们当前感兴趣的话题。识别这些主题有很重要的意义。

概率话题模型可以用于探索和预测离散数据底层结构。如 Hofmann 提出的 PLSI（Probabilistic Latent Semantic Indexing）模型[5]是通过话题的隐变量产生相关词语的生成模型。考虑由不同话题混合组成的一个文件，该模型能够通过隐变量或话题集合在一个文件中生成这些词语。将此过程反向进行，即通过观察到的数据（词语和文件）推断隐变量，从而得到基本话题分布。

## 2.3.2　OLDA 模型原理

本节主要介绍的是 LDA 的在线版本 OLDA 模型，OLDA 模型可以自动捕捉主题模式及它们随时间的变化过程。LDA 的在线工作方式能够在新的文件出现时，随即建立最新、最先进的模型（每个文件中的话题分布和每个话题中的词语分布）。因此，可以得出一种基于经验贝叶斯的解决方案，这样做能够保证在不访问之前处理过的文件的情况下，根据数据的动态变化来逐步调整模型从而得到话题分布。根据已经得到的话题分布取得对新文件的词语抽样，从而实现此方案。

OLDA 模型也用到了 Dirichlet 分布的共轭性质以保持模型结构的简单且连续。此外，OLDA 模型会在某个时间处理数据的一个子集，从而提高内存的使用效率和减少运行时间。这种动态方法用于在文本流中检测新话题并跟踪它们随时间的变化，计算根据时间排列的主题之间的相似性，还可以实时检测称为孤立点的异常话题。

## 2.3.3　OLDA 模型的先进性

OLDA 模型可以很容易地抓取话题随时间的演变。在预测未知文件时，OLDA

模型的表现可以和 LDA 相媲美，甚至超过 LDA。

本节介绍的 OLDA 模型对于离散数据可以模拟数据流中话题的时间演化，通过对文件的处理得到有意义的话题。在特定时间点检测一个小文件集合中的话题时，OLDA 的表现也优于 LDA。

下面通过一个实验来说明 OLDA 的时间效率优越性。数据集为 NIPS（神经信息处理系统）会议中 1988—2000 年的文章和路透社的 12 902 个文件。

OLDA 像标准 LDA 一样可以通过分类等方法来识别 NIPS 中有意义的话题。在路透社的每个流文件中，发现的话题都能够与文章类别很好地吻合。然而，OLDA 不需要访问整个数据集来发现这些话题，每次只需运行一部分文件，这样在时间效率和空间效率上都有很大提升。图 2-6 展示了对于路透社的数据分别执行标准 LDA 和 OLDA 所需的时间。由此图可以看出，OLDA 的执行时间是根据数据流的大小而产生的，是一个恒定时间；而执行 LDA 所需的时间是根据所分析数据的累积大小决定的。此外，LDA 需要将整个数据存储以供将来处理，但是，OLDA 模型只需存储整个数据的一部分。

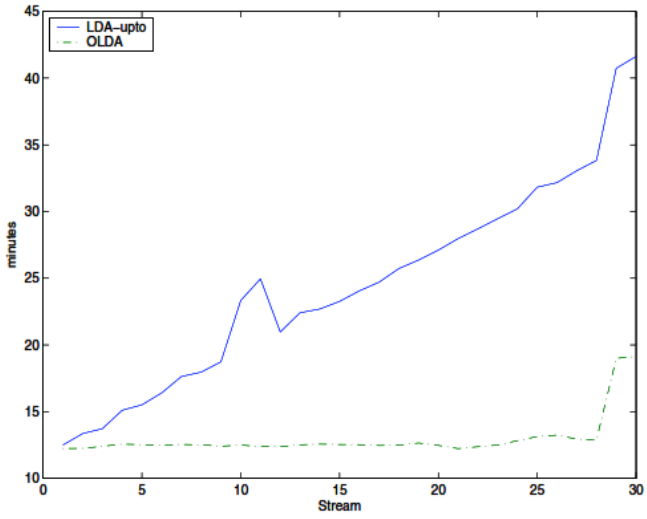



图 2-6　OLDA 和 LDA 运行时间比较

## 2.4　时间和社会话题评估（TSTE）

本节将会介绍一项新型的话题检测技术——时间和社会话题评估（TSTE）[6]，它能够实时检索社会活动中最新出现的热门话题。这个方法首先需要提取博文中的内容，并且通过一个新型衰退理论对词条的生命周期进行建模，从而挖掘新的话题内容。如果一个话题过去比较少见，但在某个特定时间段内经常出现，那么这个话题就是最新的。此外，内容的重要性还取决于此内容的来源，通过 PageRank 算法可以分析出网络中的社会关系以确定用户的影响力。最后，通过可连通话题图，能够在用户特定的时间限制内检测出新话题。

### 2.4.1　Twitter 下的 TSTE 模型简介

Twitter 作为一个社交网络平台，人们为了不同的目的在其上发表博文，包括日常的聊天、会议、分享消息或链接、发布新闻等。Twitter 最近在从社交平台向信息平台转型，并且改变了输入提示框的内容，从“你在做什么”变为“发生了什么事”。由此可见，Twitter 定义了一个低级的信息新闻发布入口。即使这个系统不能严格代表一个权威的信息媒体，但它拥有庞大的用户数量和高效的响应时间，经常能够比其他媒体更快地提供新消息。由于专业记者需要时间整合材料，专业的信息媒体总需要一定时间才能对新闻做出反应。而一个普通的 Twitter 用户不需要考虑读者的写作风格及水平，可以最快地发布新消息，这种特性使得其实时性领先于各大媒体。

下面介绍一种提取新话题的方法，它主要通过实时分析网络社区中新出现的博文来实现。总体想法是，一个新话题可以定义为在一个时间间隔内兴起但过去没有被广泛传播的话题。此过程可以基于以下 5 步来执行。

- 提取用户通过博文（包括所有语言）生成的内容并以带有相对频率的词条向量形式正规化。

- 基于用户的社会关系定义活跃用户的有向图，并通过 Page-Rank 算法计算它们的权威性。
- 对于每一个词条，根据影响用户权威性的新型衰退理论来模拟其生命周期，以便研究其在一个特定时间间隔内的使用情况。
- 通过对词条的生命状态（通过能量值定义）排序，选取一组新型术语。
- 最后创建一幅可连通的话题图，连接了提取出的词条和它们同时发生的相关词条，从而获得新的话题集合。

在本系统中，一个话题的定义是表达一个论点的一组具有连贯性的语义相关术语，如文献[7-9]，因此，对于每个时间间隔，系统能够检索出所有讨论话题中最相关的新话题。

### 2.4.2 内容提取

在大多数信息检索（IR）系统中，分析过程始于在文档流中对相关关键字的实时提取。在上文中已经提到，对 Twitter 中新话题的提取是在给定时间段内进行的，因此在一个给定时间范围 $r$ 内，定义第 $t$ 个时间间隔为

$$I^t=<i_t,i_t+r> \tag{1}$$

$i_t$ 是第 $t$ 个时间段的开始节点（$i_0$ 代表第一个实例）。在时间段 $I^t$ 内提取语料主题 $TW^t$，$n=|TW^t|$；并且为每个博文 $tw_j$ 提供一个相关向量 $t\vec{w}_j$，$t\vec{w}_j$ 将提取的相关信息规范化。为了在网络上从世界范围内快速提取到相关新闻，选择不根据其产生的语言或地区来区分消息。这种做法显然不利于对所有包括停顿词、错别字及无关词的关键词的处理。可以通过考虑逆频技术的适应性标准文本分析方法识别噪声。

通过这个方法，不仅可以保留所有关键词，还可以突出一些与话题高度相关但是出现频率不高的词。因此，可以通过下式计算第 $j$ 条博文中的第 $x$ 条相关信

息的权重 $w_{j,x}$：

$$w_{j,x}=0.5+0.5\cdot\frac{tf_{j,x}}{tf_j^{\max}} \tag{2}$$

$tf_{j,x}$ 是第 $j$ 条博文中第 $x$ 个词条的频率，$tf_j^{\max}$ 是其中最高的词条频率。

因此，对于每个博文 $tw_j$，Twitter 向量为

$$\vec{tw_j}=\{w_{j,1},w_{j,2},\cdots,w_{j,v}\} \tag{3}$$

$K^t$ 为时间段 $I^t$ 内的语料词表（关键词集合），$v=|K^t|$。

## 2.4.3　用户权威

Twitter 中的用户能够关注其他用户，被关注用户无须反向关注。因此，社交网络图为有向图 $G(V,E)$，$V$ 为用户节点，$E$ 为有向边，$<u_i,u_j>$边在当且仅当 $u_i$ 被 $u_j$ 关注时存在。

因此，需要测量 $G$ 中每个用户的重要程度，可以认为被关注量大（输入边多）的用户是社区中具有影响力的信息源。例如图 2-7 中“algore”是一个颇具权威性的用户，因为他的每条消息能立即被数以千计的用户看到。此外，用户的权威性也可以根据其追随者的重要程度来度量。“algore”具有权威性，因此被其关注的人也基于这个权威性关系相应地被假定为有更高的权威性。例如，虽然关注“current”的人不多，但是由于“algore”的关注，“current”的权威度也相对较高。可以参考 Page-Rank 算法，该算法从整个网页的拓扑图来计算节点权威。对于节点用户的权威性，计算方法如下：

$$\text{auth}(v_i)=d\times\sum_{v_j\in\text{follower}(v_i)}\frac{\text{auth}(v_j)}{\left|\text{follower}(v_j)\right|}+(1-d) \tag{4}$$

$d \in (0,1)$，follower($v_i$)为关注该节点用户 $v_i$ 的用户集合，follower($v_j$)为被该节点用户 $v_j$ 关注的用户集合。用迭代算法计算权威值，初始值为

$$\text{auth}^0(v_i)\frac{1}{|U|} \tag{5}$$

每一步算法为

$$\text{auth}^t(v_i) = d \times \sum_{v_j \in \text{follower}(v_i)} \frac{\text{auth}^{t-1}(v_j)}{|\text{follower}(v_j)|} + (1-d) \tag{6}$$

当满足收敛条件时结束处理。

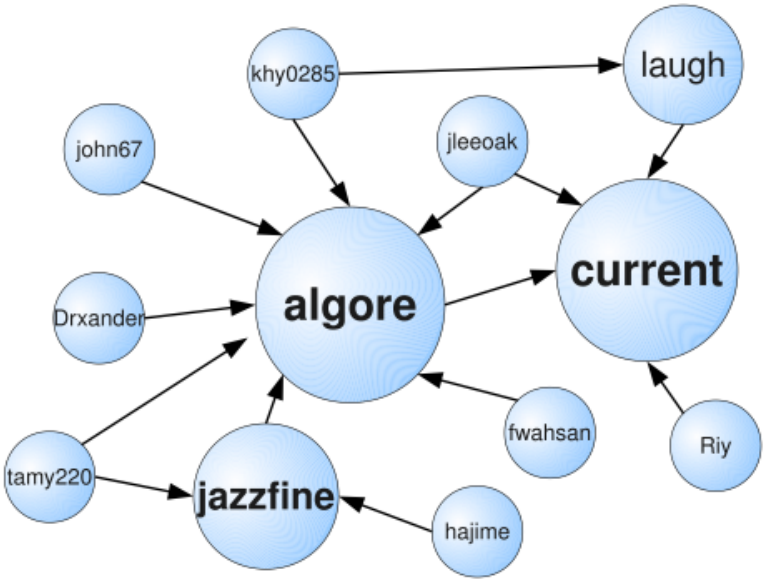



图 2-7　权威性计算：“algore”社区的样本，节点大小代表其重要性

## 2.4.4　内容衰退理论

一般来说，一个新出现的关键词可以看作近期发生的事情的相关语义单元。

为了获取这样的过滤信息，需要一个关于作者和博文时间顺序的精确模型。由于许多传统的分类和聚类策略会忽略博文中的时间关联，所以许多传统分类聚类策略并不适用。下面将每个消息短语看作一个有机体，通过分析消息短语的生命周期，以便在其成为重要消息时将其提取。

一个关键词的生命周期可以被当作一个生命体：具有丰富的营养（存在于大量相关博文）则生命周期就会被延长，营养不充足时就会慢慢死亡。

依托这个比喻，可以通过这个关键词的能量对其使用情况进行评估。它的能量表明了其生命力状态，并且可以评定关键词使用情况，具有高能量的词具有更重要的地位。

### 2.4.5　从新关键词到新话题

在本节所构建的系统中，一个话题是由一组与新兴关键词语义相关的词条组成的最小集合。因此，为了检索新兴话题，需要在特定时间间隔 $I'$ 内考察所有的博文集合，并且通过其中共同出现的信息来分析关键词的语义相关性。

下面考虑在一组博文中的关键词“victory”。单一的关键词难以确定一个话题，因此，考虑在 2008 年 11 月这个时间片段内的关键词，通过与其他关键字相关联可以轻易联想出话题，如“elections”、“USA”、“Obama”和“McCain”等。

## 2.5　话题预测分析

在 Twitter 中，有很多正在讨论的事件或话题。有些时候，人们想要知道哪些话题将会成为热门话题及为什么会成为热门话题。因此，需要预测话题趋势并且给出话题趋势变化的解释。对于 Twitter 上的话题预测，可以利用 Gerald Appel 提出的股票技术分析指数 MACD[10]。

### 2.5.1 趋势预测

将关注重点从话题的长期演化上转移到话题的短期趋势变化上，即预测在接下来几个小时内 Twitter 上的话题是否会流行起来，这种预测是实时的。

对股票和期货价格走势的趋势预测研究是较为成熟的，许多指标能够用来帮助准确地预测价格走势，如 EAM（指数移动平均值）[11]、MACD（移动平均值收敛-发散）[12]、ROC（变化率）[13]等，其中 EAM 和 MACD 是最简单并且应用最广泛的指标。

趋势动力与两个移动平均值相关。首先定义移动平均值，将连续时间分成几个连续的大小相等的时间片段，在时间片段 $t_i$ 内，关键词的出现被标注为 $f(t_i)$，移动平均值的时间窗口大小为 $k$，在第 $n$ 个时间片段，移动平均值 MA($n$,$k$)为：

$$\mathrm{MA}(n,k)=\frac{\sum_{i=n-k+1}^{n} f(t_i)}{k} \tag{7}$$

如果 $n<k$，则移动平均值为

$$\mathrm{MA}(n,k)=\frac{\sum_{i=1}^{n} f(t_i)}{k} \tag{8}$$

移动平均值能够很好地追踪话题趋势，并且不同大小的值能够追踪不同大小时间片段内的趋势，正如原始 MACD，趋势动力定义如下：

$$\mathrm{TM}(n)=\mathrm{MA}(n,k_s)-\mathrm{MA}(n,k_l) \tag{9}$$

$k_s$ 是较短的时间窗口大小，$k_l$ 是较长的时间窗口大小，趋势动力是用短移动平均值减去长移动平均值。然而在 Twitter 上话题特征与股票略有不同，当一个关键字所代表话题的长期移动平均值在一段时间内保持较高水平时，这个话题会维

持长时间的热度。但是若一只股票的价格在一段时间保持较高，则它可能只是降价的开始，因此重新定义趋势动力：

$$\mathrm{TM}(n)=\mathrm{MA}(n,k_s)-\mathrm{MA}(n,k_l)^{\alpha},0<\alpha<1 \tag{10}$$

趋势动力和原始 MACD 之间有两个不同之处：一是使用了移动平均值（MA）而非指数移动平均值（EMA）；二是由于关键词的频率在不同的时间片内的表现很不稳定。EMA 更多地取决于最近时间片段上的频率，它的波动性同样较大。所以使用 MA，它的表现更为平滑。另外，可以给较长的 MA 提供折扣参数，尽管有些较热门的话题与冷门话题具有相似的趋势动量，更热门的话题会具有更大的长周期 MA，因此向其提供折扣参数$\alpha$。

然而公式（10）的趋势动力值不稳定，因此在实践中会使用移动平均值将其平滑化为

$$\mathrm{Momentum}(n)=\mathrm{MA}(\mathrm{TM}(n),k) \tag{11}$$

在对股票的 MACD 分析中，有一些对股票未来价格趋势预测的规则，可以选择最简单有效的规则。若该话题趋势动力值由负变为正，则该话题趋势将上升；若其由正变为负，则话题将会衰退。

## 2.5.2　趋势变化的原因

趋势改变的原因主要有三个方面。

（1）关键用户：当一些有影响力的用户参与某些新闻话题的讨论时，此话题可能会变得更加热门，因此需要找出当话题变热门时哪个具有影响力的用户参与了此话题的讨论。将话题变热门时的时间片段表示为 $T$，新话题为 $TP$，可以得到表 2-1 所示的统计信息。

表 2-1　在时间片段 $T$ 内的用户数量和主题 $TP$ 的统计

| | $TP$ | $\overline{TP}$ |
|---|---|---|
| $T$ | $A$ | $B$ |
| $\overline{T}$ | $C$ | $D$ |

- $A$ 表示在时间片段 $T$ 内发布了博文并讨论话题 $TP$ 的用户数量。
- $B$ 表示在时间片段 $T$ 内发表了博文但是没有讨论话题 $TP$ 的用户数量。
- $C$ 表示在时间片段 $T$ 内没有发表博文但是讨论了话题 $TP$ 的用户数量。
- $D$ 表示既没有在时间片段 $T$ 内发布博文又没有讨论话题 $TP$ 的用户数量。

之后可以计算在时间片段 $T$ 内用户 $U$ 和话题 $TP$ 的相关度，利用相关度为用户排序，并且选出排名靠前的用户。

（2）关键词：一个新话题变得更热门的原因可以是与之相关的话题变得热门，如果使用关键词来表示其他事件或者话题，那么这个问题变为在话题变热门的时间片段内寻找最相关的关键词。可以先通过计算得出在时间片段 $T$ 内词语 $W$ 和新话题 $TP$ 之间的相关性，然后对相关度排序并选出排名靠前的关键词。

（3）话题互动：一个新话题的衰落是由于讨论人数的减少，当话题热度上升时，它会吸引越来越多的关注。用户会将关注点放在上升趋势的话题上，因此，其他话题的讨论度就会相对减少，所以话题间存在相互影响。这是 Twitter 上话题衰落的一个重要因素。

## 2.6　异常检测算法下的话题发现

上述方法都是基于在消息交换中词条的频率来检测从而发现新话题的，然而这不适用于现今的社交网路。因为社交网络用户越来越倾向于发表非文本内容，如网址图片或视频等。为了处理非文本内容，Takahashi 等人[14]提出从用户之间动态有意或无意产生的链接中发现突发性话题，而不是只关注文本内容。在社交网

络上，用户之间可以通过回复，提及“@”或转发来相互交流。“@”也是其中一个很重要的行为，他们提出了一个社交网络用户“@行为”的概率模型，并产生了一个算法，此算法结合了“@”异常分数和基于 SDNML 的 Changepoin 检测技术。异常分数是通过概率模型针对每个用户的标准行为来计算的，并通过此模型执行异常检测来检测新话题的出现。

## 2.6.1　概率模型简介

“@”行为可以通过不同的形式将同一个社交网络中的用户相连接，如回复、转发或直接在文本中@某用户。一篇帖子可能会包含一些@信息，有些名人用户可能会不停地收到@消息，普通人被@的情况可能很少。

本节主要通过检测用户的@行为来检测新话题的出现。我们的基本假设是：一个新兴的话题是人们评论或转发较多的话题。由于同义词或同音词的原因，传统的基于词频的方法效果可能会不好。同时，传统方法需要根据目标语言进行复杂的预处理（如分割），也不能在非文本信息中很好地应用。而通过@所得到的词是唯一的，仅需要少量的预处理工作即可获取，因为在文本中它通常会被分离出来。

下面描述的概率模型可以抓取用户的正常@行为，包括每篇帖子包含的@数目和@到的用户频率。该模型可以定量测量此帖子发出后对用户的影响。这种技术可以检测统计相关结构下的变化，并且查明该主题在哪里出现。

## 2.6.2　概率模型方法

模型整体流程图如图 2-8 所示，假设通过一些 API 获取一系列社交网络服务中的行为数据信息。对过去时间间隔 $T$ 内响应用户的每篇新帖子进行抽样，训练下面的@模型。基于学习到的概率分布来给每篇帖子分配异常分数，然后将分数聚合，以便在后续的分析中使用。

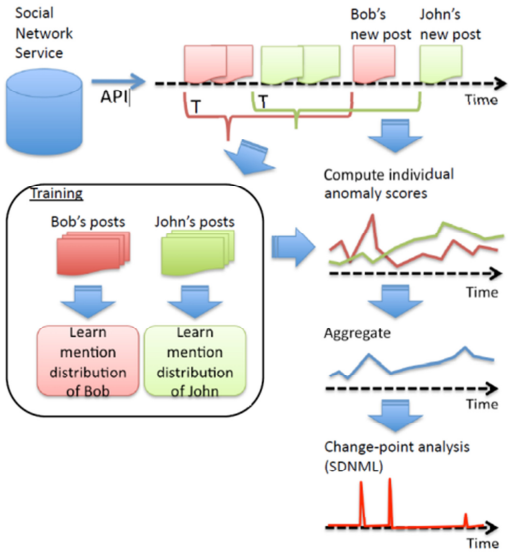



图 2-8　概率模型流程图

概率模型：帖子包含的@的数量 $k$ 和用户 ID 集合 $V$ 可以作为社交网络流的识别标志，通常会考虑下述概率分布：

$$P(k,v|\theta,\{\pi_v\}) = P(k|\theta)\prod_{v\in V}\pi_v \tag{12}$$

这里的联合分布由两部分组成：@的数量 $k$ 的概率和给出@数量后每个@的概率。@数量概率 $P(k|\theta)$定义如下：

$$P(k|\theta)=(1-\theta)^k\theta \tag{13}$$

另外，在 $V$ 中提及用户的概率为独立同分布，假设有 $n$ 个训练样本 $\Gamma=\{(k_1,V_1),\cdots,(k_n,V_n)\}$概率分布 $P(k,V|\Gamma)$为：

$$P(k,V|\Gamma)=P(k|\Gamma)\Pi_{V\in v}P(V|\Gamma) \tag{14}$$

首先计算对于@数量的预期分布，可以假设一个 Beta 分布作为先验分布。先验分布的密度函数如下：

$$p=(\theta|\alpha,\beta)=\frac{(1-\theta)^{\beta-1}\theta^{\alpha-1}}{\mathrm{B}(\alpha,\beta)} \tag{15}$$

$\alpha$和$\beta$是 Beta 分布的参数，B($\alpha$,$\beta$)是 Beta 函数。通过贝叶斯公式，预期分布可以通过下式获得：

$$\begin{aligned}P(k\mid\Gamma,a,b)&=\frac{P(k,k_1,\cdots,k_n|\alpha,\beta)}{P(k_1,\cdots,k_n|\alpha,\beta)}\\&=\frac{\int_0^1(1-\theta)^{\sum_{i=1}^n k_i+k+\beta-1}\theta^{n+1+\alpha-1}\mathrm{d}\theta}{\int_0^1(1-\theta)^{\sum_{i=1}^n k_i+\beta-1}\theta^{n+\alpha-1}\mathrm{d}\theta}\end{aligned} \tag{16}$$

分子和分母的积分都能通过 Beta 函数获得，预期分布可以重新表示为：

$$P(k|\Gamma,\alpha,\beta)=\frac{\mathrm{B}(n+1+\alpha,\sum_{i=1}^n k_i+k+\beta)}{\mathrm{B}(n+\alpha,\sum_{i=1}^n k_i+\beta)} \tag{17}$$

$m=\sum_{i=1}^n k_i$ 是训练集合 $\Gamma$ 中的@总数。

下一步，派生@用户 $v$ 的预期分布，定义最大似然估计（ML）$P(v|\Gamma)=m_v/m$，$m$ 为@总数，$m_v$ 为在数据集 $\Gamma$ 中用户 $v$ 的@总数。然而，ML 依然不能解决不能出现在Γ中的用户，它会将这些用户的概率全部置零，这会使得本节所述框架产生异常。因此，可以使用 CPR[15]，它维持了没有在 $\Gamma$ 中@的用户概率比例 $\gamma$。已知用户概率如下：

$$P(v|\Gamma)=\frac{m_v}{m+\gamma}\ (m_v\geqslant 1) \tag{18}$$

另外，新用户的@概率如下：

$$P(\{v{:}m_v=0|\Gamma\}=\frac{\gamma}{m+\gamma} \tag{19}$$

计算链路异常分数：对于用户 $V$，为了计算在时间 $t$ 包含 $k$ 个@的用户 $u$，计

算新帖子的异常得分 $x=(t,u,k,V)$，需要在训练集 $\Gamma_u^{(t)}$ 上计算概率（17），之后链路异常得分为：

$$s(x)=-\log P\left(k \mid \Gamma_u^{(t)}\right)-\sum_{v\in V}\log P(k\left|\Gamma_u^{(t)}\right.) \tag{20}$$

上面两式可以通过计算@数量的预期分布得出。

结合不同用户的异常分数：公式（20）中的异常得分是根据用户当前所发帖子和过去行为计算的，为了衡量用户行为总趋势，可以利用离散窗口 $\tau>0$ 汇总得到异常分数：

$$s_j^{'}=\frac{1}{\tau}\sum_{t_i\in[\tau(j-i),\tau j]}s(x_i) \tag{21}$$

$x_i=(t_i,u_i,k_i,V_i)$是用户 $u_i$ 在时间 $t_i$ 给用户 $V_i$ 发的包含 $k_i$ 个@的帖子。

## 2.7 本章小结

下面对本章内容进行总结，表 2-2 在以下方面对算法进行了比较：

（1）如何定义话题。

（2）该方法可以处理哪些内容。

（3）该方法是检测热门话题还是预测热门话题。

表 2-2 话题检测总结

| 方　法 | 话题定义 | | | 内容类型 | | 任务类型 | |
|---|---|---|---|---|---|---|---|
| | 单条短语 | 短语集合 | 分布式 | 文本 | 非文本 | 检测 | 预测 |
| PT | √ | | | √ | | √ | |
| OLDA | | | √ | √ | | √ | |
| TSTE | | √ | | √ | | √ | |

续表

| 方　法 | 话题定义 | | | 内容类型 | | 任务类型 | |
|---|---|---|---|---|---|---|---|
| | 单条短语 | 短语集合 | 分布式 | 文本 | 非文本 | 检测 | 预测 |
| SDNML | √ | | | √ | √ | √ | |
| MACD | √ | | | √ | | | √ |

虽然话题检测技术已有很多方法，但是还有很多挑战需要解决，如数据复杂性等问题，将来用户会更多地使用如图像、视频或网页链接等其他形式的数据，这会增加数据的复杂性，而现今的大多数方法只能解决社交网络中的文本内容。

# 参考文献

[1] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In CSCW ’11, (short paper) 2011: 355-358.

[2] G. Salton and M. McGill. Introduction to modern information retrieval. McGraw-Hill, Inc. New York, NY, USA, 1986.

[3] L. AlSumait, D. Barbar´a, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In ICDM ’08, 2008: 3-12.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” The Journal of Machine Learning Research, 2003, 3: 993-1022.

[5] T. Hofmann, “Probablistic Latent Semantic Indexing,” Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, 1999.

[6] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In MDMKDD ’10, 2010:4-13.

[7] G. P. C. Fung, J. X. Yu, H. Liu, and P. S. Yu. Time-dependent event hierarchy construction. In KDD'07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 300-309, New York, NY, USA, 2007. ACM.

[8] Allan, James, ed. Topic detection and tracking: event-based information organization. Vol. 12. Springer Science & Business Media, 2012.

[9] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Simple semantics in topic detection and tracking. Inf. Retr., 2004, 7(3-4):347-368.

[10] L. Rong and Y. Qing. Trends analysis of news topics on Twitter. International Journal of Machine Learning and Computing, 2012, 2(3):327-332.

[11] A. J. Lawrance and P.A.W. Lewis, "The exponential autoregressive-moving average earma (p, q) process," Journal of the Royal Statistical Society, Series B (Methodological), 1980: 150-161.

[12] G. Appel, "Become Your Own Technical Analyst," The Journal of Wealth Management, 2003, 6(1): 27-36.

[13] G. Appel and F. Hitschler. Stock Market Trading Systems. Traders Press, 1990.

[14] T. Takahashi, R. Tomioka, and K. Yamanishi. Discovering emerging topics in social streams via link anomaly detection. In ICDM '11, pages 1230-1235, 2011.

[15] D. Aldous, "Exchangeability and related topics," in Ecole d'Ete de Probabilites de Saint-Flour XIII—1983. Springer, 1985: 1-198.

# 第 3 章　影响力最大化

## 3.1　引言

社交网络描述了群体中人与人之间的关系和交互作用，在信息的传播中起到至关重要的媒介作用。若想加强一个话题在社交网络中的传播程度，则需要找到社交网络中最具影响力的节点，并且了解每个节点在被周围的节点影响到何种程度时会被同化。近些年，关于社交网络用户间相互影响的研究受到了广泛关注，商业领域中的口碑营销是其重要应用场景，这种营销策略与病毒传播相似，因此被称为“病毒营销”[1-5]。如何实现影响力最大化在社会生产中有着重要的研究价值，应用领域有农业生产、公益活动、舆情分析、产品营销等。例如，在产品营销中，一家公司需要找到一个由少部分人组成的集合，希望这些人能够在市场中发挥最大的影响力，从而推动产品的销售。最终目的是通过影响力最大化方法，

找到最优的种子节点集合来激活其周围的节点，使得最终被激活的节点数达到最大值。

关于如何在特定的网络传播模型下找到这样一组节点，目前已有很多研究，但由于网络数据量的不断增长，很多算法在大型网络中计算代价过高，或者难以处理动态网络结构，因此一些更加先进的算法仍在不断出现。

本章主要内容安排如下：第二节首先分析社交网络中影响力最大化的基本模型和概念；第三节至第七节分别介绍影响力最大化问题的基本算法和优化算法，例如，新鲜度衰减下的影响力最大化算法（IMND）及社交网络信息覆盖最大化（MCIP）等；第八节对影响力最大化问题所面临的挑战和发展前景进行展望。

## 3.2 影响力最大化基本概念

### 3.2.1 影响力最大化的描述

在线性阈值和独立级联模型中，影响力最大化的解能够达到（$1-\frac{1}{e^{-\varepsilon}}$）近似最优（e 是自然对数的底数，$\varepsilon$ 是任意正实数），贪心爬山策略可以实现这种近似，并且始终能保证至少在最优值的 63%。这些内容可以利用子模函数理论来证明。下面介绍子模函数。

**子模函数：**对于任意函数 $f(\cdot)$，将一个无限集合 $U$ 映射到一个非负实数，当 $f(\cdot)$符合“收益递减”属性时称其为子模函数：即当 $S \subseteq T$ 时，将一个元素添加到集合 $S$ 中的边际收益至少相当于将一个元素添加到集合 $T$ 中的边际收益，表示如下：

$$f(S\cup\{v\})-f(S)\geqslant f(T\cup\{v\})-f(T) \tag{1}$$

$S$、$T$ 满足 $S \subseteq T$。下面介绍影响力最大化问题的目标。

首先选择一个初始节点集合 $A$，线性阈值模型和独立级联模型（以及模型的一般化推广）都会以一组初始活跃节点 $A$ 展开扩散，定义 $A$ 的影响范围为 $\sigma(A)$，$\sigma(A)$表示在 $A$ 的影响下，最终活跃节点个数的期望值。给出一个参数 $k$，影响力最大化问题需要找到一个 $k$ 个节点的集合，使得最终影响范围最大。

影响力最大化问题属于 NP-hard 问题，其中，NP 是指非确定性多项式（Non-deterministic Polynomial）。所谓的非确定性是指，可用一定数量的运算去解决多项式时间内可解决的问题。多项式时间在计算复杂度理论中指的是一个问题的计算时间 $m(n)$不大于问题大小 $n$ 的多项式倍数。NP 问题通俗来说是其解的正确性能够被“很容易检查”的问题，这里“很容易检查”指的是存在一个多项式检查算法对其正确性做出判断。相应地，若 NP 中的所有问题到某一个问题都是图灵可归约的，则该问题为 NP-hard 问题。到目前为止，这类问题中没有一个找到有效算法。倾向于接受这类问题不存在有效算法这一猜想，认为这类问题的大型实例不能用精确算法求解，必须寻求有效的近似算法。对于独立级联模型和线性阈值模型，影响力函数 $\sigma(\cdot)$为子模函数，影响力最大化问题是 NP-hard 问题。

## 3.2.2　社交网络的马尔科夫模型

考虑 $n$ 个潜在消费者，令 $X_i$ 为第 $i$ 个消费者（$X_i$ 为布尔变量），如果第 $i$ 个消费者购买了市场中的产品，则 $X_i$ 为 1，否则为 0。$X_i$ 的邻居节点为 $N_i=\{X_{i,1},\cdots,X_{i,ni}\} \subseteq X-\{X_i\}$，$X_i=\{X_1,\cdots,X_n\}$，$X_i$ 与 $X\text{-}N_i\text{-}\{X_i\}$之间相互独立。$X^k(X^u)$是已知（未知）值的节点，令 $N_i^u(M)=N_i \cap X^u$。假设产品由一系列特征 $Y=\{Y_1,\cdots,Y_m\}$ 来描述，如果 $M_i$ 为布尔变量，对于第 $i$ 个消费者，$M_i=1$ 表示其享受了折扣，反之 $M_i=0$，$M_i$ 也可以表示第 $i$ 个消费者享受的折扣值。$M=\{M_1,\cdots,M_n\}$，对于任意 $X_i \notin X^k$，有

$$
\begin{aligned}
&P\left(X_i \middle| X^k, Y, M\right) \\
&= \sum_{c(N_i^u)} P(X_i, N_i^u \middle| X^k, Y, M) \\
&= \sum_{c(N_i^u)} P(X_i \middle| N_i^u, X^k, Y, M) P(N_i^u \middle| X^k, Y, M) \\
&= \sum_{c(N_i^u)} P(X_i \middle| N_i^u, Y, M) P(N_i^u \middle| X^k, Y, M)
\end{aligned} \tag{2}
$$

上述模型为社交网络下的马尔科夫市场模型。

如果 $M$ 是一个布尔变量，令 $c$ 为市场对于一个消费者的投入成本（$c$ 为常数）$r_0$ 为在没有市场行为（例如折扣）的情况下将产品销售给某个消费者的收益，$r_1$ 为有市场行为的情况下的收益。$r_0$ 和 $r_1$ 只有在市场行为为折扣的情况下不等，其余情况下都相等。令 $f_i^1(M)$表示将 $M_i$ 设置为 1，其余为 0，$f_i^0(M)$相似。不考虑对其他消费者的影响，市场对于消费者 $i$ 的预期利润提升为：

$$
\begin{aligned}
&\mathrm{ELP}_i(X^k, Y, M) = \\
&r_1 P(X_i = 1 \middle| X^k, Y, f_i^1(M)) \\
&-r_0 P(X_i = 1 \middle| X^k, Y, f_i^0(M)) - c
\end{aligned} \tag{3}
$$

令 $M_0$ 为全零向量，对于所有消费者的预期利润提升为：

$$
\begin{aligned}
&\mathrm{ELP}(X^k, Y, M) = \\
&\sum_{i=1}^{n} r_i P(X_i = 1 \middle| X^k, Y, M) - \\
&r_0 \sum_{i=1}^{n} P(X_i = 1 \middle| X^k, Y, M_0) - \left|M\right| c
\end{aligned} \tag{4}
$$

如果 $M_i$=1，则 $r_i$=$r_1$；反之，如果 $M_i$=0，则 $r_i$=$r_0$。|$M$|是 $M$ 中值为 1 的数量。最终目标是找到 $M$ 的值的分布，使得 ELP 最大。通常来说，找到最优的 $M$ 需要尝试所有情况，由于这种困难性，提出以下几种近似过程。

**单程路径（Single Pass）：**对于每一个消费者 $i$，如果 $\mathrm{ELP}(X^k, Y, f_i^0(M))>0$，则令 $M_i$=1，否则 $M_i$=0。

**贪心查找（Greedy Search）**：令 $M=M_0$，循环查找 $M_i$，如果 $\mathrm{ELP}(X^k,Y,f_i^1(M))>\mathrm{ELP}(X^k,Y,M)$，则将 $M_i$ 置为 1，直到不存在 $i$ 使得 $M_i=1$。

**爬山查找（Hill-climbing Search）**：令 $M=M_0$，当 $i_1=\mathrm{argmax}_i\{\mathrm{ELP}(X^k,Y,f_i^1(M))\}$ 时，$M_{i1}=1$；当 $i_2=\mathrm{argmax}_i\{\mathrm{ELP}(X^k,Y,(f_i^1(M)))\}$时，$M_{i2}=1$。执行此方法直到不存在 $i$ 使得 $M_i=1$。

上述方法的计算花销依次增大，但精确度同样依次增大，从而可以不断优化 ELP 值。

## 3.3　影响力最大化基本算法

### 3.3.1　启发式算法

启发式算法是相对于最优化算法提出的，一般是根据人们的直观印象或者根据以往的经验而构造的算法，它在一定可接受的时间和空间代价的范围内，给出了组合优化问题的一个可行解。对于影响力最大化问题，经常用到的启发式算法有如下几种。

（1）基于 High-Degree 的启发式算法：依照节点的度数来进行节点的选择，是常见的启发式算法之一。在无向图中节点的度数为其邻居节点的个数。High-Degree 算法按照节点的度数从大到小进行排序，选择前 $k$ 个节点作为初始节点。

（2）基于 Distance-Centrality 的启发式算法：将节点之间的路径长度作为评价尺度。该算法按照节点与其他节点之间平均距离递增的顺序，选择前 $k$ 个节点作为初始节点。

（3）Random 算法：随机选择 $k$ 个节点作为传播的初始节点。

### 3.3.2 贪心算法

贪心算法的主要思想是：我们在解答某个问题时，需要做出判断和决定，而每次做出的选择在当时的情况下来看都是最好的，从某种意义上来说，贪心策略求解的结果可以看作局部的最优解。我们在设计贪心算法时，关键在于如何选择贪心策略，因为贪心算法多数情况下求得的只是某个局部的最优解。另外，在设计贪心算法时，在某个时刻所做出的选择只与当前的状态相关，而之后发生的过程并不会影响当前所做的选择和状态。

贪心算法选择初始节点的过程如下：

（1）首先将初始集合 $S_0$ 置为空集，即 $S_0=\emptyset$。

（2）记 $f(S_i)$为初始集合为 $S_i$ 时，最后处于活跃状态的节点总数。

（3）设 $g(u|S_i)=f(S_i\cup\{u\})-f(S_i)$，$g(u|Si)$表示将节点 $u$ 添加到集合 $S_i$ 中所得到的边际收益。在迭代过程中，每次向集合中添加边际收益最大的节点，即 $u_i = \text{argmax}_{v\in V-S_{i-1}} g(v|S_{i\text{-}1})$，令 $S_i=S_{i-1}\cup\{u_i\}$。

（4）重复上面的过程，直到迭代 $k$ 步为止，得到的 $S_k$ 即为贪心算法大小为 $k$ 的初始集合。

贪心算法的伪代码表示如下：

```
算法 1   Greedy(G,k)
输入：社会网络图 G；初始集合大小 k
输出：初始集合 S
1 初始化 S=NULL；
2 for i=1 to k
3    for each vertex v in set V-S
4    计算节点 v 的边际收益 g(v|S)=f(S∪{v})−f(S)
5       选出边际收益最大的节点 u=argmax_{v∈V−S} g(v|S)
6    end for
```

```
7    令 S=S U{u}
8 end for
9 输出集合 S
```

## 3.4　新鲜度衰减情况下影响力最大化算法

影响力最大化在实际生活中得到了广泛应用，但是若一个信息反复出现在用户面前，则这个信息的影响力会逐渐减弱。例如，在 Twitter 中，一个用户更可能会转发第一次阅读到的某条消息，而随着这条消息出现频率的增高，用户对其转发的可能性也会降低，这种现象被称为新鲜度衰减。但是影响力最大化算法大多数没有考虑新鲜度衰减对信息传播的影响。本节讨论新鲜度衰减下影响力最大化的问题。图 3-1 为新鲜度衰减的一个例子。

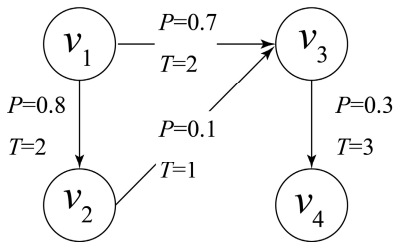



图 3-1　在有向边上影响概率为 $P$、时延为 $T$ 的社交网络有向图

4 个相邻用户之间的有向边表示他们之间的影响关系，每一条有向边上的两个值分别为影响概率 $P$ 和时延 $T$。例如，$v_1$ 能够以 0.7 的概率影响 $v_3$，时延为 2 个时间单位。给出种子节点集合$\{v_1,v_2\}$，在不考虑新鲜度衰减的情况下，节点 $v_3$ 能够被激活的概率为 0.1+(1−0.1)×0.7，0.1 是 $v_3$ 被 $v_2$ 激活的概率，(1−0.1)×0.7 是 $v_3$ 被 $v_1$ 激活的概率。如果 $v_1$ 和 $v_2$ 激活 $v_3$ 的顺序是一定的，那么在影响力传播过程中就需要考虑到新鲜度衰减现象。因此，如果 $v_2$ 首先尝试激活 $v_3$，那么 $v_3$ 被 $v_1$ 激活的概率将会减弱［少于(1−0.1)×0.7］。

不同于传统的影响力传播模型，这种影响力函数不是单调或子模函数。由于每一对节点之间的时延不同，影响力计算方法会变得较为复杂。下面提出的算法可以建立一种传播路径来估计种子节点的影响力传播范围，这种算法可以高效地实现大范围传播。

### 3.4.1 新鲜度衰减函数

对于某个事件，假设一个节点被 $n$ 个活跃邻居节点尝试激活了 $n$ 次，令 $TP_n$ 为一个节点被 $n$ 个节点尝试激活后被成功激活的概率，$p_n$ 为一个节点在被第 $n$ 个节点尝试激活后被成功激活的概率。下式为 $TP_n$ 和 $TP_{n-1}$ 之间的关系模型：

$$TP_n=TP_{n-1}+(1-TP_{n-1})\times p_n \tag{5}$$

之后计算 $p_n$：

$$p_n=(TP_n-TP_{n-1})/(1-TP_{n-1}) \tag{6}$$

为了将新鲜度衰减函数 $f(n)$正规化，令 $f(n)=p_n/p_{n-1}$，$p_1=TP_1$，$p_1$ 是所有节点在第一次被尝试激活后就成功被激活的概率。新鲜度衰减函数正规化之后的形式为 $f(n)=\Upsilon n-1$，采用最小二乘法估计参数 $\Upsilon$。

### 3.4.2 独立级联模型下的新鲜度衰减

考虑独立级联模型情况（表示为 $IC_{ND}$），每个节点都有两个状态：活跃和非活跃。节点可以由活跃状态变为非活跃状态，但是不能由非活跃状态变为活跃状态。每个节点都有两个参数：影响概率 $P_{uv}$ 和影响时延 $T_{uv}$。

给出一幅有向图 $G=(V,E)$，集合 $S\subseteq V$，一个新鲜度衰减函数 $f(n)$，$IC_{ND}$ 模型的流程如下：在 $t(t\geqslant 0)$时刻，活跃的节点集合为 $A_t$，令 $A_0=S$。每个节点 $u\in A_t$ 在 $t+T_{uv}$ 时间点都有一次机会尝试激活其邻居节点 $v$，成功激活节点 $v$ 的概率为

$P_{uv}\times f(n)$，$n$ 表示 $v$ 被尝试激活的次数。过程进行直到没有任何尝试激活的行动，所有活跃节点的集合为 $\sigma(S)=\sum_{t=0}^{\infty}\left|A_t\right|$。

$IC_{ND}$ 模型下的影响力最大化问题为 IMND（Influence Maximization with Novelty Decay）问题。

### 3.4.3　贪心算法的优化

由于在 $IC_{ND}$ 模型中影响力函数既不是单调函数也不是子模函数，传统的贪心算法变得不再适用，因此下面介绍一种新算法。

#### 1．R-Greedy 算法

IMND 问题需要找到一组节点 $S$，$|S|\leqslant K$。最初选择 $K$ 个节点，其中的每个节点都能使得边际影响力最大化，之后选出能够使得影响力最大化的节点集合 $S$。这种算法被称为局限性贪心算法（Plane Restricted Greedy Algorithm），即 R-Greedy 算法。尽管返回的一组节点个数可能小于 $K$，但由于有限的预支，这种情况在实际中很少出现。

#### 2．动态精简优化（DP）

令 $S_k$ 表示被选中的节点集合，R-Greddy 算法需要计算每个节点 $u\in V\backslash S_{k-1}$ 的边际影响力增益，并且在每次迭代中调用算法检查第 $k$ 个节点 $S_k$。为了更高效地实现这种算法，需要对该算法进行优化处理。下面介绍动态精简优化。在 R-Greedy 算法中利用上面介绍的影响力传播过程选择潜在种子节点，完整算法如下：

DP 下的 R-Greddy 算法

**Input**:G=(V,E),$T_{UV}$,$P_{UV}$,f(·),K

**Output:S**

1 $S\leftarrow\emptyset, S_0\leftarrow\emptyset, \sigma(S)\leftarrow 0, \sigma(S_0)\leftarrow 0, s\leftarrow$NULL;
2 对于任意 $v\in V$，计算$\sigma(\{v\})$，并且将其插入 $Q_0$ 中;
3 **for** $k\leftarrow 1$ **to** K **do**
4 maxMarInf$\leftarrow -\infty$;
5 **for** node $u\in V\backslash S_{k-1}, \sigma(\{v\})\geq$maxMarInf **do**
6 **If** $u\in Q_{k-1}$,and
$(\inf_{k-1}^{u}+\sigma(\{S_{k-1}\})-\sigma(S_{k-1})<\max$ MarInf
Then
7 继续
8 **Else**
9 计算 $\inf_k^u$ 并且将$(u, \inf_k^u)$插入 $Q_k$ 中;
10 **If** $\inf_k^u-\sigma(S_{k-1})>$ maxMarInf **then**
11 maxMarInf$\leftarrow \inf_k^u-\sigma(S_{k-1})$;
12 $s_k\leftarrow u$;
13 $S_k\leftarrow S_{k-1}\cup\{s_k\}$
14 $\sigma(S_k)\leftarrow\sigma(S_{k-1})+$maxMarInf
15 S←从 k=1 到 k=K 中使得$\sigma(S_k)$最大的 $S_k$
16 return S

令 $Q_k$ 存储已经检查过的候选节点，$inf_k^u=\sigma((S_{k-1})\cup\{u\})$代表在向节点集合 $S_{k-1}$ 中添加节点 $u$ 之后的影响力。在每一轮的循环中，DP 检查所有的候选节点，一旦节点的影响力低于阈值就立刻结束。maxMarInf 记录了在第 $k$ 个循环中最大的边际收益。另一个最主要的优化是，如果 $u$ 在第（$k$–1）轮循环中被检查过（第 5 行），就能推导出它的边际影响力上限($inf_{k-1}^u+\sigma(\{S_{k-1}\})-\sigma(S_{k-1})$（第 6 行），如果上限大于 maxMarInf，则节点被忽略（第 7 行）。此算法计算出了 $S_{k-1}\cup\{u\}$的影响力，并且将其存储在 $Q_k$ 中（第 9 行）。如果边际收益大于 maxMarInf，则更新 maxMarInf 和 $S_k$（第 10～12 行）。最后获得了种子节点 $S_k$ 及其影响范围$\sigma(S_k)$。

DP 优化删除了影响力小于 maxMarInf 的节点，并且这个算法依然保持着 R-Greedy 算法的特性。

### 3.4.4　影响力传播计算算法

在优化了 R-Greedy 算法后，下一个问题就是计算种子节点的影响范围。为了解决这个问题，需要找到种子节点的传播路径。

**新鲜度衰减模型下的传播路径**

给出一组种子节点集合 $S\subseteq V$，最终影响范围为 $\sigma(S)=\sum_{u\in V}AP_S(u)$，$AP_S(u)$ 为 $u$ 被 $S$ 激活的概率。为了计算 $AP_S(u)$，需要定义新鲜度衰减模型下的传播路径：给出一个种子节点集合 $S$ 和一幅有向图 $G=\{V,E\}$，当且仅当在 $k>1$ 的情况下，当 $i\neq 1$，$u_1\in S$ 并且 $u_k\notin S$ 时，图 $G$ 中的路径 $h=(u_1\xrightarrow{e_1}u_2\xrightarrow{e_2}u_3\cdots\xrightarrow{e_{k-1}}u_k)$ 为一条新鲜度衰减下的传播路径（$\mathrm{PP_{ND}}$）。

由于一个节点不能被多次激活，所以一条 $\mathrm{PP_{ND}}$ 路径不能包含重复节点，当影响概率为 $\prod_{i=1}^{i=k-1}P(e_i)\times\hat{E}(\tau^h(u_{i+1}))$ 时，路径长度为 $\mathrm{Len}(h)=\sum_{i=1}^{i=k-1}T_{e_i}$，$\hat{E}(\tau^h(u_{i+1}))$ 为对 $u_{i+1}$ 新鲜度衰减值的期望。

下一步需要计算 $\hat{E}(\tau^h(u))$。一条 $\mathrm{PP_{ND}}$ 路径只有两种状态：如果一条路径上的所有节点 $u_1, u_2, u_3, \cdots, u_k$ 都为活跃状态，则这条路径为连通状态；否则为阻塞状态。如果 $h_c$ 是 $\mathrm{PP_{ND}}$ 路径中第 $c$ 短的路径，那么就存在 $c-1$ 条更短的路径，想要计算 $\hat{E}(\tau^{h_c}(u))$，就需要考虑到所有 $c-1$ 条路径是连通的还是阻塞的。例如，如果 $h_1$ 和 $h_2$ 是最短的两条路径，那么想要计算 $\hat{E}(\tau^{h_3}(u))$ 就需要考虑 4 种情况。

下一步就是找到 $\mathrm{PP_{ND}}$ 路径。对于一个给定的种子节点集合 $S$，用 $\mathrm{PP_{ND}}(u,S)$ 表示从 $S$ 到 $u$ 的所有 $\mathrm{PP_{ND}}$ 路径。若集合 $S$ 过大，则会使得 $\mathrm{PP_{ND}}$ 路径的数量过大，找到 $\mathrm{PP_{ND}}(u,S)$ 的计算量也过大。为了解决这个问题，首先可以删除概率小于阈值

$\theta$（$\theta$>0）的路径，其次可以去除路径长度大于 $c$ 的路径（因为节点被激活次数过多后不易被激活），$PP_{ND}$ 被表示为 $PP_{ND\theta,C}(u,S)$。

计算 $PP_{ND\theta,C}(u,S)$之后，最后一步是计算$\sigma(S)$，所有以 $u$ 为终点的路径都执行上述过程，最终得到$\sigma(S)$。

## 3.5 社交网络中信息覆盖最大化

前两节提出的社交网络影响力最大化研究，其目的是选择一组种子节点，使得传播结束后获得的活跃节点数最多，达到影响范围最大化。但是最后的活跃节点数并不能完全代表真实的信息覆盖情况。一种常见的情况是，当某个节点的活跃邻居节点尝试激活它时，即使没有激活成功，该节点也会得到该消息，故称之为消息节点。因此，当研究社交网络中信息覆盖最大化时，除了需要考虑活跃节点的数量，消息节点的影响也不能忽略，进而产生了信息覆盖最大化的问题，这个问题的目的是找到最多的活跃节点和消息节点。

### 3.5.1 信息覆盖最大化问题简介

现有的很多模型都在描述消息的传播过程，例如独立级联模型[5]、线性阈值模型[6]、基于数据的信用分部模型[7]和线性社交影响模型[8]等。在这些模型中，独立级联模型和线性阈值模型是随机扩散模型。上述模型中的节点只存在两种状态：活跃和非活跃。活跃节点可以被视为接受了这条消息并会再次传播这条消息的节点；而非活跃节点则是未被激活且不会传播这条消息的节点。

然而，在实际传播中非活跃节点也存在两种类型。例如，对于一条发布在 Twitter 上的消息，有些用户会转发这条消息，有些用户则不会，但是在没有转发的用户中，有些人由于邻居节点的转发而得知了这条消息成为消息节点，其余节

点为真正的非活跃节点。

可以推断，一个节点想要成为消息节点，那么它的邻居节点中至少有一个为活跃状态；相反，如果一个节点的全部邻居节点都为非活跃状态，那么此节点永远不可能成为消息节点。现实世界中消息节点的数量非常庞大，但是由于影响力最大化问题只考虑活跃节点而忽略了消息节点，所以它不能很好地模拟现实世界。为了更好地衡量信息覆盖范围，应将这两类不会再次传播消息的节点都列入考虑范围。

因此，可以考虑从非活跃节点中找出消息节点，并发掘消息节点的价值，以更好地衡量信息覆盖范围，新的问题就演变为最大化活跃节点和消息节点的集合问题。

## 3.5.2　信息覆盖最大化问题的特征

独立级联模型是解决此问题的最佳模型，因此，以下分析全部基于独立级联模型。在此模型中，选出的种子节点会传播某条消息并且试图激活周围的邻居节点，如果某个节点被尝试激活，那么它会成为消息节点或者活跃节点，活跃节点继续尝试激活其邻居节点，过程进行直到没有节点被激活。令 $S$ 为种子节点集合，$A$ 为活跃节点集合，$L$ 为消息节点集合，此问题的公式化描述如下[9]：

$$\operatorname{argmax}_S F(S)=E(|A|)+E(|L|) \tag{7}$$

$$\text{s.t.}|S|=k$$

$k$ 为种子节点数量的预期值，$E(\cdot)$为活跃节点集合或消息节点集合的预期值，$F(S)$为活跃节点和消息节点的预期值。

在现实世界中，消息节点和活跃节点的价值是不一样的，因此，引入权重来调整消息节点对信息覆盖最大化的贡献，描述如下：

$$\text{argmax}_S F(S)=E(|A|)+\lambda E(|L|) \tag{8}$$

$$\text{s.t.}|S|=k,\lambda=[0,1]$$

$\lambda$ 是权重系数，当 $\lambda$=0 时，此问题退化为一般的影响力最大化问题；当 $\lambda$=1 时，消息节点具有和活跃节点同等的重要性。

下面介绍信息覆盖最大化问题的几个特征。

特征 1：对于独立级联模型下的信息传播网络，信息覆盖最大化问题是 NP-hard 问题。

特征 2：对于独立级联模型下的信息传播网络，加权信息覆盖最大化问题是 NP-hard 问题。

特征 3：对于独立级联模型下的信息传播网络，$F(\cdot)$为单调子模函数。

特征 4：对于独立级联模型下的信息传播网络，$W(\cdot)$为单调子模函数。

### 3.5.3 信息覆盖最大化问题的解决方法

解决信息覆盖最大化问题有两种方法。第一种为“懒惰前进”贪心算法。由于该问题具有子模性质，相比普通的贪心算法，“懒惰前进”贪心算法能够有效地减少计算 $F(\cdot)$或者 $W(\cdot)$的时间。虽然“懒惰前进”贪心算法显著降低了时间成本，但是仍然难以适应大规模的网络。为了解决这个问题，需要不断更新算法。

第二种算法为“基于度的启发式算法”。重新回顾目标函数，可以看到一个节点对网络的贡献程度依赖于其出度，因此，如果根据出度将节点排名，并取前 $k$ 个节点作为种子节点，则可以得到一个好的结果。更进一步，当一个节点被选中时，与其出边相连的邻居节点会成为消息节点，这些消息节点有可能是其他节点的出边邻居节点，从而导致其他节点的“有效出度”降低。这种现象意味着可以动态调整每个节点的“有效出度”，使得算法更加高效。

## 3.6　在线影响力最大化

对于现有的影响力最大化算法，假设两个节点 $a$ 和 $b$ 之间存在从 $a$ 指向 $b$ 的一条边，那么 $a$ 对 $b$ 的激活概率为给定值 $p$。然而，这种假设并非适用于所有情况。考虑一个刚搬到某城市的商场，这个商场能得知一些用户在社交网络上的信息，但是不能得知这些用户之间的影响如何传播[10]。在这种情况下，只有得知影响概率才能执行影响力最大化算法并找到种子节点集合。得到这类信息的办法可以是使用“行为日志”来记录社交网络用户过去的行为，不过一般来讲日志记录无法在短期内获得。

尽管在社交网络中影响概率未知，但依然可以执行影响力最大化算法，这个问题称为在线影响力最大化问题，目标是在得到影响概率的同时执行影响力最大化算法。

### 3.6.1　在线影响力最大化问题描述

在缺乏影响概率信息的情况下，如何找到使得影响力最大的种子节点集合是一个很大的挑战。为了解决这个问题，提出以下解决方案：对种子节点执行多轮选择，在每一轮中，激活一些选出的种子节点（例如，发放免费商品使得种子节点用户推荐给其邻居节点用户），根据这些节点的反馈来决定下一轮选择激活哪些种子节点，根据每一轮的执行逐渐学习影响概率。

图 3-2 为在线影响力最大化问题框架，它包含了多个相互影响的进程，每个进程完成一个或两个目标：（1）激活影响力大的节点；（2）进程执行过程中学习影响概率。进程包含两个阶段：选择和行动。在选择阶段需要维护好不确定概率分布图，这幅图建立了社交网络用户之间不确定的影响概率分布。基于现有的影

响力最大化解决方案，在这幅图中执行种子节点选择策略，最终产生 $k$ 个种子节点。在行动阶段，将选中的节点在现实世界中激活（向选中用户投放广告），这些用户的行动（或反馈）被用于更新影响概率分布图。迭代继续进行，直到市场预算枯竭。下面从两个阶段的角度分别进行分析[14]。

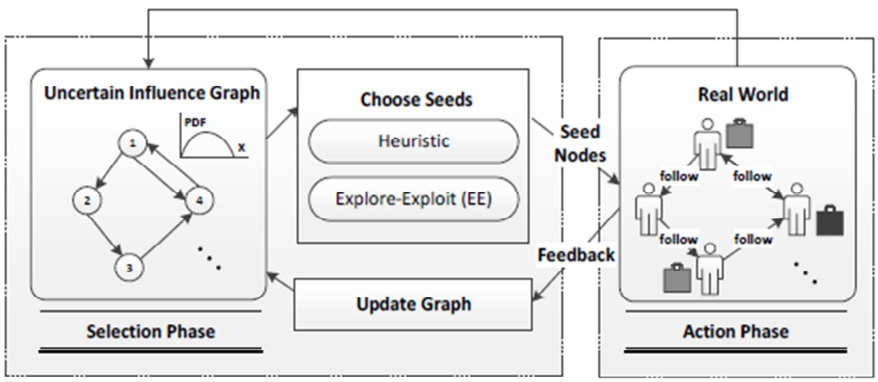



图 3-2　在线影响力最大化框架

## 3.6.2　节点选择策略

要在执行过程中选择种子节点，一种最简单的方法就是利用现有的影响力最大化算法。但是由于缺乏对影响概率的了解，这种方法可能不是最好的。因此，另一种称为 Explore-Exploit（EE）的方法产生了，这种方法根据现有的影响概率执行影响力最大化算法。

Exploit：从影响概率图中选出对影响力传播最有价值的 $k$ 个种子节点，可以利用任何最先进的影响力最大化算法（例如，CELF[11]、DD[12]、TIM、TIM+[13]）。

Explore：通过一些策略选择 $k$ 个种子节点来提高对影响概率的认知。

在影响概率图下，相比使用一般影响力最大化算法，对 EE 算法的合理使用会有更好的表现。

在在线影响力最大化解决方法中，有 $N$ 个进程需要执行，在每个进程中都需要执行现有的影响力最大化算法，如果 $N$ 太大，就会对算法的表现有所影响。如果底层的影响概率图很大，那么这种影响会更严重，运行时间会被严重加长。存在一种有效的解决办法，可以令用户的反馈只影响一小部分影响概率图。

### 3.6.3　更新不确定影响概率图

可以看出，种子节点的选择策略在图 3-2 中所示的影响概率图中执行。这幅图需要精确反映现在所能得知的节点之间的相互影响，因此，节点选择策略是现有的最优策略。下面执行基于活跃节点反馈的算法来更新影响概率图。此过程基于两个变量进行研究，两个变量为对影响概率图的更新是局部还是全局。在影响概率图中，局部更新会更新两个节点之间的概率参数，而全局更新会更新整个影响概率图的概率参数。这种算法是基于经典机器学习理论而开发的（例如，最小二乘法和最大似然法），并且这种算法具有广泛的应用，可以解决很多实际问题（例如，当影响图底层概率未知的情况下向市场推广一款商品）。

## 3.7　流式子图的增量算法

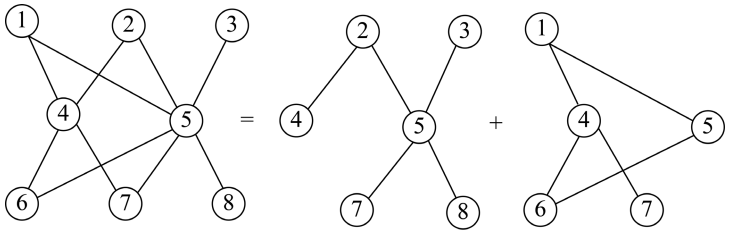
前面几节介绍了很多社交网络影响力最大化算法，但是在具有数以百万计节点的网络中，这些算法依然难以执行。现有的启发式算法和贪婪算法的计算量都很大。所以，本节对解决大规模网络中影响力最大化问题提出一种新的算法。首先将大规模网络划分为小的子图，再依次对每张子图执行影响力传播分析。该算法称为增量算法。

### 3.7.1 大规模网络下影响力最大化问题

影响力最大化问题可以看作独立级联模型和线性阈值模型下的离散优化问题。它已经被证实为 NP-hard 问题，通过贪心算法可以达到近似最优。然而随着社交网络的规模不断增加，贪心算法正逐渐变得不适用。

为了在大规模网络下解决影响力最大化问题，最直观的想法是使用启发式算法，但是这种算法在大规模网络结构下并不能很好地执行。另外，可以利用先进的贪婪算法来加速节点选择过程[8,12,14]。在选择新的节点时，这些方法着重于减少不必要的计算。然而，这些算法仍然难以处理数以百万计的大规模网络节点。更重要的是，所有这些贪心算法都难以处理现实世界中的动态网络。下面提出的增量算法估计了种子节点的预期影响力传播范围，并对其进行评估。具体来说，这种算法将大图划分为小型子图，将子图作为数据流进行处理，然后将对子图处理之后的结果重新组合，恢复至整个网络中。这种增量过程也能够处理随时间变化的动态网络。

这个问题最主要的困难是小的子图之间非独立。例如，在图 3-3 中，右边的两张子图共同享有节点 4,5,7，更严重的情况下两张子图甚至共享一条边。在这种情况下，我们假设不同子图之间的边是不相交的，每张子图被转换为一类强连通分量（SCCs）。




（a）Original Graph　　（b）Subgraph 1　　（c）Subgraph 2

图 3-3　大规模网络分解图示

### 3.7.2　增量算法的特征

上文提出的增量算法的特征总结如下：

（1）一张网络大图被分解为子图，并且每张子图可以并行处理。

（2）每张子图被转换为一类强连通分量（SCCs）。

（3）此算法显著降低了整个过程的时间复杂性。

## 3.8　线性阈值模型下的可扩展社交网络影响力最大化

### 3.8.1　问题描述

独立级联（IC）和线性阈值模型（LT）刻画了社交网络的两个不同方面，IC 模型侧重于个人[1]（独立的）交互和社交网络中朋友之间的相互作用，而 LT 模型侧重于影响力传播过程中的阈值行为。我们在生活中经常可以看到，当一个人的周围有足够多的朋友都在玩同一款网络游戏时，他（她）也会被带动玩同一款游戏。这两种模型的广义推广模型是等价的，但是基本的线性阈值模型和独立级联模型是两种不同的模型。

我们在第一节中也提到过，两种模型下的影响力最大化问题都是 NP-hard 问题，并且提出了一种贪心算法能够达到 63%的近似最优值。然而贪心算法的效率并不令人满意。为了克服原贪心算法的低效率，有很多人提出了贪心算法的优化或者提出了新的启发式算法[12,13,15,16]。然而以往提出的各种启发式算法都是基于独立级联模型来设计的，对于同样重要的线性阈值模型却大都不适用，并且对于线性阈值模型并没有适用的可扩展启发式算法。

### 3.8.2 LDAG 算法

下面提出一个为线性阈值模型量身定制的可扩展影响力最大化算法，称为 LDAG 算法。首先构造一个本地 DAG，将节点 $v$ 的影响力限制到本地 DAG 结构中，这使得在小的 DAG 中影响力计算变得易于处理并且快速。关于节点 $v$，该算法向本地 DAG 中逐步添加节点使得这些节点对 $v$ 的单独影响大于阈值参数 $\theta$。这种本地 DAG 结构使得 LDAG 算法非常高效。

当社交网络为有向无环图（DAG）时，影响力计算的时间复杂度与图的大小呈线性关系。这种算法可以扩展到数百万的节点和边，比一般的贪心算法速度更快，并且此算法专为线性阈值而设计，在实际场景中的表现非常稳定。因此，此算法是线性阈值模型下的最佳算法。

## 3.9 本章小结

随着社交网络成为一种商业平台，基于社交网络的影响力最大化研究已经成为最具有研究价值的领域。本章介绍了几种影响力最大化传播模型和优化算法，从而不断提升算法性能。

研究社交网络中的影响力最大化问题，需要整合计算机学科及经济学等其他学科的核心研究领域，具有很强的社会价值和商业价值。

## 参考文献

[1] David Kempe, Jon Kleinberg, Eva Tardos. Maximizing the Spread of Influence through a Social Network. KDD, 2003.

[2] F Bass. A new product growth model for consumer durables. Management Science 15(1969), 215-227.

[3] J Brown, P Reinegen. Social ties and word-of-mouth referral behavior. Journal of Consumer Research 14:3(1987), 350-362.

[4] P Domingos, M Richardson. Mining the Network Value of Customers. Seventh International Conference on Knowledge Discovery and Data Mining, 2001.

[5] J Goldenberg, B Libai, E Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. Marketing Letters 12:3(2001): 211-223.

[6] M Granovetter. Threshold models of collective behavior. American Journal of Sociology , 1978, 83(6):1420-1443.

[7] Amit Goyal, Francesco Bonchi, Laks VS Lakshmanan. A data-based approach to social influence maximization. Proceedings of the VLDB Endowment, 2011, 5(1): 73-84.

[8] Biao Xiang, Qi Liu, Enhong Chen, Hui Xiong, Yi Zheng, Yu Yang. Pagerank with priors: An influence propagation perspective. IJCAI, 2013.

[9] Maximizing the Coverage of Information Propagation in Social Networks. Physics and Society, 2015.

[10] M J Lovett, R Peres, and R Shachar. On brands andword of mouth. J Marketing Research, 2013, 50(4).

[11] J Leskovec, A Krause, C Guestrin, C Faloutsos, J VanBriesen, N Glance. Cost-effective outbreak detection in networks. KDD 2007.

[12] W Chen, Y Wang, and S Yang. Efficient influence maximization in social networks. In KDD, 2009.

[13] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. KDD, 2010: 1029-1038.

[14] Y Tang, X Xiao, and Y Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. SIGMOD 2014.

[15] M Kimura and K Saito. Tractable models for information diffusion in social networks. European Conference on Machine Learning and Knowledge Discovery in Databases, 2006: 259-271.

[16] J Leskovec, A Krause, C Guestrin, C Faloutsos, J VanBriesen, N S Glance. Cost-effective outbreak detection in networks. Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2007: 420-429.

# 第 4 章　收益最大化

## 4.1　引言

在对影响力最大化问题进行了深入探讨之后，我们接下来进一步讨论收益最大化问题，这是一个更加实际的问题。社交网络的出现让我们第一次有了数以百万计的好友，其中的信息对于广告、个性化推荐和促进商业智能化具有很大的价值。然而，现实世界中社交网络的潜在价值和其表现出的实际价值之间还存在着显著的差异。

由于其蕴藏的巨大潜在价值，因而产生了大量对社交网络盈利性的研究工作[1-6]，这些工作主要是为了设计高效的营销策略。将商品出售给一个买家往往会给其他买家带来一些影响，这种效应被称为交易的外部性，能够促进下一次销售并产生

收入的性质称为正外部性。在现实生活中，许多人会由于他的朋友购买了某种商品而决定购买此种商品，因此利用正外部性能够最大限度地提高卖方的收益。

在一个营销方案中，如病毒营销或其他方式，产品定价和用户对其估值两个因素共同决定了用户是否购买一个商品。一个人对产品的估值为其愿意付出的最高金额，若不愿购买，则估值为 0[16]。在这种情况下，卖方需要找到一种能够使其产品广泛被采用的营销策略。

## 4.2 最佳营销策略模型

### 4.2.1 模型简介

在最佳营销策略模型[1]中，买方是否要购买某个商品是由该商品的价格和其他购买该商品的消费者对其影响所共同决定的。

若每个买家的情况完全相同，则可以在多项式时间内找到最佳营销策略。在一般情况下，我们都可以研究出这个问题的近似算法。这种策略被称为影响-拓展策略。此策略首先将商品免费赠予一些人，从而通过他们来进一步影响其他买家；之后利用“贪婪”价格策略从剩余买家的购买中提取收益。

随着社交网络的普及，一些公司得以收集用户的个人信息及其社会关系信息。现有社交网络大多是通过广告来赚钱[7-9]的，而本节提出的模型主要通过智能化的销售策略建立盈利社交网络。

### 4.2.2 正外部性

下面站在卖方角度介绍正外部性的例子。

（1）商品的信息往往依靠口碑来传播。例如，我们可能会因为朋友购买了某个商品而得知或者购买这个商品或服务。当我们的朋友购买了某种商品的其中一个时，我们可以通过评估其质量从而决定是否购买。一个高质量的商品会促使我们去购买，甚至不惜花更高的价钱。

（2）有时候，商品会有明确的能够促进传播的特性。例如，很多音乐播放器都会有一个音乐分享功能，允许用户以无线方式与其他人交换音乐。

一个有远见的卖家可以利用正外部性来增加收入。例如，卖家为了增加商品销量，最初可以免费提供商品给一些受欢迎的买家。事实上，这种策略早已应用于实践。

关于免费提供试用品的基本思想可以用以下几种方法来概括：

（1）卖家可以对商品实施折扣销售而非完全的免费提供。这里需要对折扣进行权衡，更大的折扣可能意味着交易中单个商品获得的收益减少，但同时也会增加商品销量及其在未来购买者中的影响力。那么，折扣应该多大呢？

（2）由于影响力往往是不对称的，销售行为发生的顺序会对外部效果产生影响，通常，高知名度及拥有广泛社交关系的用户会有更大的影响力。因此，若一种销售行为具有促进进一步销售的潜力，卖方会希望此销售行为能够尽早发生。所以，应该采取怎样的销售顺序呢？

本节的目的就是探讨使得卖家的收入达到最优的营销策略。

### 4.2.3　模型结果

下面从数字商品的销售行为来研究收益最大化营销策略。生产一个数字商品的成本为零。假设有一个潜在的买家集合 $V$，买方是否购买一个商品是由其他拥有此商品的买家意见和此商品本身的价格决定的，这个模型模拟了已经购买商品

的买家对其他买家的影响。尽管卖方不知道具体的影响程度，但是却可以知道有关买方分布的信息。一般来说，降低价格会使得销量增加。

在营销策略中，卖方以某种序列将买家排列，并且对每个买家提供一个价格。当买方接受这个价格后，卖方得到该商品的收益。营销策略有两个要素：卖方给买方提供的商品序列和商品价格。一般而言，最有利的方法是让序列中有影响力的买家尽可能早地购买商品。为了达到这个目标，卖家甚至可以向他们提供更低的价格促使他们购买该商品。

（1）对称设置：对于卖家而言，所有买家在购买商品前的身份都是相同的，在这样的设置中，可以忽略买家序列，使用动态规划法得出最佳定价政策。最佳营销策略表现为以下行为：买方接受卖方报价的概率随着营销策略的进展而越来越低。最初，最佳营销策略提供商品折扣来促使买方购买商品。对于此商品而言，这种行为会增加购买序列中买家的数量值，这使得最优策略可以从后续买方中得到更多收入。在现实生活中，最优策略甚至在早期序列中赠送免费商品。

（2）常规设置：接下来，考虑常规设置下的最佳营销策略算法。首先表明，寻找最佳营销策略是 NP-hard 问题，因此我们可以考虑使用近似算法。

在此我们提出一种简单的营销策略：影响-拓展策略，此策略在后面的章节中会有详细的介绍。回想一下，任何营销策略都有两个方面：定价和寻找买家顺序。影响-拓展策略分为两步：第一步，影响，通过在对称形式下最佳营销策略的激励，卖方选择一个集合 $A \subseteq V$，对此集合中的买方提供免费商品；第二步，拓展，卖方以一个随机序列访问剩余买家（$V \setminus A$），并通过提供最优价格来使得最终收益最大化。需要注意，这里忽略了集合 $V \setminus A$ 中买家的相互影响（集合 $A$ 中的买家类似于社交网络中的意见领袖[10]）。

如果收入函数为子模函数，则集合 $A$ 的预期收益同样为子模函数。但是由于收益函数并非单调函数，因此不能使用由 Nemhauser、Wolsey 和 Fisher 提出的简单贪婪策略[11]。

### 4.2.4　市场策略

正如上文中所讨论的，由于买方相互影响，卖方可以精心制定销售序列，并提供智能化的折扣，从而优化其收入。下面我们介绍可行销售策略空间的定义。

一个营销策略中的卖方会以某种序列访问每个买家，并且为每个买家提供一个价格，买家可以接受这个价格并支付或者拒绝购买此商品。我们假设每个买家仅会被访问一次。不论是所提供的价格还是访问顺序都是自适应的，例如，价格的设定和访问顺序可以基于此买家对商品接受或拒绝的历史。因此，营销策略会根据历史函数来寻找下一个进行访问的买家及对其提供的价格。买方只能通过已经购买了某产品的买方的影响决定是否购买该产品。在任何时间点，如果买方集合 $S$ 购买了某商品，则买方 $i$ 的值是 $v_i(S)$。

营销策略的运行包括一个价格序列，每个价格对应一个买家，价格设定由此买家接受或者拒绝其他商品的历史行为决定。运行收益是所有接受商品的买方提供的价格总和，我们称能达到最佳收益的营销策略为最佳营销策略。

### 4.2.5　对称设置最佳营销策略

假设买方值是根据对称设置定义出来，$|V|$个买方中每个买方个体的值都服从一个分布 $F_k$。

现在可以推导出最优营销策略。由于买方是对称设置，所以对买方的访问顺序是可以忽略的。此外，价格函数仅与已购买者数量和未购买者数量有关。若 $k$ 个人已经购买了某商品，$t$ 个人还没有购买该商品（包括正在考虑是否购买的买方），令 $p(k,t)$为卖方根据最佳营销策略提供给处于考虑状态买方的价格，$R(k,t)$为从剩余买方中期望得到的最大收益。现在给定一个价格 $p$，如果买方接受，那么可以得到收益 $p+R(k+1,t-1)$；如果买方拒绝，则收益为 $R(k,t-1)$。那么买方当且仅当自己的心理价位大于等于 $p$ 时接受此商品，概率为 $1-F_k(p)$。最终设置能够使得

预期剩余收益实现最大化的 $p$，对于任意 $p$，预期剩余收益为：

$$F_k(p)\cdot R(k,t-1)+(1-F_k(p))\cdot(R(k+1,t-1)+p) \quad (1)$$

在上式中对 $p$ 进行微分，使微分后的值为 0 则可以得到使收益最大化的 $p$ 值。

$$F_k(p)\cdot(R(k,t-1)-R(k+1,t-1)-p)+1-F_k(p)=0 \quad (2)$$

之后我们可以设定满足上述公式的 $p(k,t)$值，变量 $R(k,t)$的值便能很容易计算出来了。上述动态程序可以在购买者数量平方的时间内解决，收益最低的情况为 $R(k,t)=0$。至此定义了最佳营销策略，所需仅是密度函数，并无其他额外假设。

下面进行总结：假设 $S\subseteq V\backslash\{i\}$，$i$ 的值在$[0,|S|+1]$上均匀分布，$F_k$ 的值在$[0,k+1]$上均匀分布。图 4-1 和图 4-2 描绘出最佳价格随着 $k$ 和 $v$ 的变化趋势。图 4-1 证实，对于一个固定的 $t$，最佳价格随着已购买者数量的增长而增长。图 4-2 证实，对于一个固定的 $k$，随着未购买者数量的增加，保证未购买者接受该商品的重要性大于每个商品的单独收益。图 4-2 还表明，在营销策略算法起步阶段，大量买方处于未购买状态，最佳价格为 0。

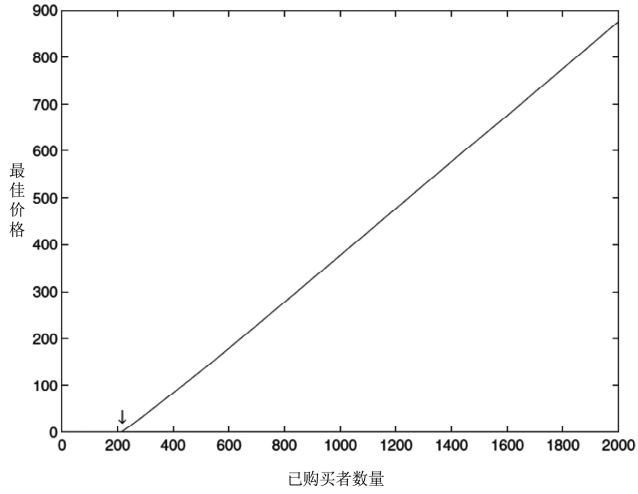



图 4-1　最佳价格随着已购买者数量变化而变化的情况

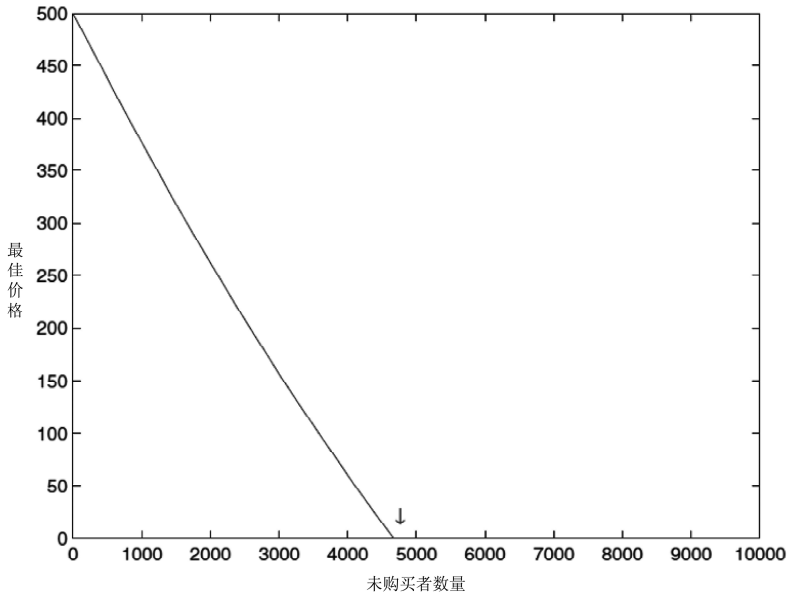



图 4-2　最佳价格随着未购买者数量变化而变化的情况

### 4.2.6　影响-拓展营销策略

上文提到过营销策略有两个要素：价格和提供商品的顺序。下面我们将详细介绍影响-拓展策略。该策略的第一步为影响，即向被选中的买家提供免费商品；第二步为拓展，主要基于一个随机序列和一个贪婪价格策略。

（1）将商品免费送给有重要地位的买家促使影响的产生。

（2）由于最佳营销策略为 NP-hard 问题，以随机序列选取节点最终可以达到 1/2 近似最优[7,8]。因此，在拓展的步骤中，卖方将会以随机序列访问买家。

（3）我们将在拓展这一步骤中采用最佳价格，从而最大限度地得到某买家的收益而不用担心对他人的影响。

简单定义影响-拓展策略，主要包含两个步骤。

（1）影响：向集合 $A$ 中的买家免费提供商品。

（2）拓展：以一个序列 $\sigma$（从所有可能的序列中的随机均匀选取）访问集合 $V\setminus A$ 中的买家，假设集合 $S\subseteq V\setminus\{i\}$为在买方 $i$ 之前已经买到商品的买方集合，提供给买方 $i$ 的最佳价格是分布为 $F_{i,S}$ 的函数。需要注意最佳价格是自适应的，并且是基于销售历史的。

虽然集合 $A$ 不会提供任何收益，但是可以保证他们使用商品后可以影响其他买方。这样可以从集合 $V\setminus A$ 中提取额外的收益，从而弥补甚至超过提供免费商品的损失。

## 4.3 影响-拓展策略的效率

### 4.3.1 营销策略的社交网络模型

在上文提出的社交网络收益最大化营销模式下，产品会以正外部性被销售给一组潜在买方，卖方会寻找一种营销策略，以某种顺序向买方提供价格并售卖，从而最大化自己的收益。下面将围绕均匀加和模型和影响-拓展策略展开研究，可以明显改善最大化收益的近似值。具体来说，在均匀加和模型中，对最佳营销策略和最佳影响-拓展策略的计算都是 NP-hard 问题。可以观察到，在影响-拓展策略中把价格压低可以使近似值得到改进。

现在采用文献[12]中的模型，产品会通过正外部性出售给一组潜在的买家。正如文献[12]中假设的，商品无限供应并且生产成本为 0，社交网络 $G$ 是由一组潜在买家组成的，网络中的一条边 $e=(j,i)$表示 $j$ 拥有某产品并且对 $i$ 有一个正面影响。$i$ 的值是由一个增函数 $v_i(S)$决定的，$S$ 为已经拥有某产品的买家并且对 $i$ 有积极影响。精确值 $v_i(S)$对于卖方是未知的，并且被视为随机变量，仅其分布 $F_{i,s}$ 已知。在文献[12]中，凹图形模型中的每一个 $v_i(S)$是关于 $S$ 的子模函数，均匀加和

模型中的非负权值 $W_{j,i}$ 与网络中的边$(j,i)$相关，$v_i(S)$在 0 与从 $S$ 到 $i$ 所有边相加得到的权重总和之间均匀分布。均匀加和模型中一个重要的特殊情况是无向网络，对于每一条边$(j,i)$，$W_{j,i}=W_{i,j}$。

在这种背景下，卖方只对每一个潜在买家进行一次访问，并向其提供一个个性化价格。营销策略决定了卖方访问买方的顺序和对其提供的价格，每个买家选择支付或者拒绝，拒绝后不会再被访问。卖方的目标是计算一个能最大限度地提高自己收益的营销策略。

### 4.3.2　影响-拓展策略的效率

尽管影响-拓展策略简单优雅并且效率高，然而它对于收益最大化的表现和它多项式时间内的近似性却不易被理解。本节集中在均匀加和模型上，通过得到一个关于影响-拓展策略的效率和近似性的综合性结果，使均匀加和模型中营销策略的近似性得到显著提升。

对于均匀加和模型而言，在无向社交网络中计算最佳营销策略和最佳影响-拓展策略都是 NP-hard 问题。在对影响-拓展策略的性能进行系统性研究之后，可以发现，如果使用一个更低的价格（价格的接受概率会增加），影响-拓展策略的效率会变好。因此，可以认为该策略的价格接受概率为 $p\in[1/2,1)$。

## 4.4　线性阈值模型下的收益最大化问题

本节扩展了传统的线性阈值模型，将价格和估值都纳为用户购买商品的决策因素。此模型下预期的收益函数在一定条件下保持子模性质，但是不再单调，不同于预期的影响力函数。为了最大限度地扩展 LT 模型下的收益，可以使用三种收益最大化算法。下面将对这三种算法进行简单介绍。

### 4.4.1 用户估值线性传播模型（LT-V）

首先我们要了解影响力最大化（$I_{NF}M_{AX}$）这一概念。在影响力最大化算法中，我们只需要考虑影响力权值和网络结构这两个因素，并且其营销策略为严格的二元决策，即对于网络中的任何节点，影响力最大化算法只能够决定买家对某种商品接受或者拒绝。为了解决上述限制，我们结合价格和估值，进而提出社交网络中收益最大化的问题（$P_{RO}M_{AX}$）。收益最大化问题主要研究在某一个传播模型下，我们能够找到一个最佳策略使得整个传播过程结束后的预期总收益达到最大。将线性阈值（LT）模型扩展产生一个新的传播模型，称为用户估值线性传播模型（LT-V），只有公司提供的价格不超过用户估值时，用户才会接受商品。

在 LT-V 模型中，社交网络用有向图 $G=(V,E)$表示，其中，每个节点 $u_i \in V$ 都与一个特定的价格估值 $v_i \in [0,1]$相关，其中 $v_i$ 独立随机地选自一些营销公司已知的价格估值分布。令 $v_i$ 的分布函数为 $F_i(x)=Pr[v_i \leqslant x]$，其密度函数为 $f_i(x)=\frac{\mathrm{d}}{\mathrm{d}x}F_i(x)$，假设商品价格和用户估值都属于[0,1]，因此这两个函数的定义域都为[0,1]。在经典 LT 模型中，每个节点 $u_i$ 都有一个随机阈值 $\theta_i \in [0,1]$，每条边$(u_i,u_j) \in E$ 都有一个权值 $w_{i,j} \in [0,1]$。对于每个节点 $u_i$，$\sum_{u_i \in N^{in}(u_j)} w_{i,j} \leqslant 1$，若$(u_i,u_j) \notin E$，则定义 $W_{i,j}=0$。根据文献[14,15]中的描述，假设每个种子用户都有一个固定购置成本（例如邮寄广告或优惠券的花费）。

图 4-3 给出了 LT-V 模型的状态图。在任何步骤，节点都处于非活跃、受影响和接受三种状态中的一种。LT-V 模型下的传播过程是在离散时间下进行的。最初，所有节点都处于非活跃状态；在 0 时刻，种子节点集合 $S$ 被选中并变为受影响状态；接下来，网络中的每个用户 $u_i$ 都会被系统分配一个价格 $p_i$，令 $\boldsymbol{p}=(p_1,\cdots,p_{|V|}) \in [0,1]^{|V|}$ 为价格向量，在传播过程中保持不变。对于任意 $u_i \in S$，如果满足 $p_i<v_i$，在 0 时刻都有且仅有一次机会选择转变为接受状态，否则一直处于受影响状态。

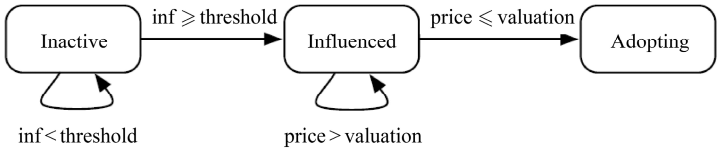



图 4-3　LT-V 模型状态图

在 $t \geqslant 1$ 之后的任意时间点，当处于非活跃状态下的节点 $u_j$ 的所有邻居节点对其影响达到 $u_j$ 本身的阈值时，$u_j$ 变为受影响状态。之后，如果 $u_j$ 在某一时刻 $t$ 满足 $p_j \leqslant v_j$，则 $u_j$ 会变为接受状态；否则将会一直保持在受影响状态。该模型是渐进的，这意味着所有节点将保持在接受或者受影响状态而不会变回非活跃状态，直到没有节点状态发生变化，则此传播过程结束。

## 4.4.2　定价策略

正如 Kleinberg 和 Leighton[13]指出的，人们在出于对卖家的信任接受一个价格之前通常不愿意透露自己的估值。此外，在得知商品价格后，由于隐私问题，他们通常只表示出自己的决定（即“是”或“否”），但并不会分享出自己的真实估值信息。因此，参考文献[14,15]，可以使用独立专用值（IPV）假设，即每个用户的估值随机独立分布于某一特定分布。这样的分布能够从一家营销公司的历史销售数据获得。此外，这个模型假设用户作为价格接受者，仅仅基于自己的估值和所提供的商品价格对这一给定价格做出一个大致的回应。

对于网络中的任何节点，收益最大化算法需要决定是否将其作为种子，并且决定应该提供怎样的价格。因此，收益最大化的目标函数（即预期总利润）是种子集合和价格向量二者的结合函数。由于需要向种子节点提供折扣，其利润函数是非单调的。此外，对于任意固定价格向量，不管估值分布的特征怎样，利润函数都会保持子模性质。

由于上述原因，收益最大化问题天生就比影响力最大化问题复杂。为了解决

此问题，我们需要设计一些更加复杂的算法。由于利润函数的形式不同于单调子模函数和线性函数，我们可以设计一种“未预算贪婪”（U-Greedy）的种子集选择框架。在每次迭代中，算法选择能达到最大边际收益的节点，直到总收益开始减少。对于任何固定的价格向量，U-Greedy 能够达到略低于（1-1/e）（其中 e 为自然常数）的近似值。为了获得完整的收益最大化算法，可以使用以下三种定价策略：ALL-OMP（最佳短视价格）、FFS（自由种子）和 PAGE（价格感知贪婪策略）。前两种定价策略是基础，它们以特定的方式选择价格，不需要考虑网络结构和影响的蔓延；而 PAGE 会在 U-Greedy 的每一轮中动态确定最优价格。实验表明，对于预期利润和运行时间，PAGE 表现最佳。

**ALL-OMP**：OMP 为最佳短视价格。社交网络中的价格接受者都是短视的，只能站在自己的角度给出对所提供价格的大致反应，即对价格的高低进行主观评价。我们给出价格估值 $v_i$ 的分布函数 $F_i$，则 OMP 便可以根据下式计算出来：

$$p_i^m = \arg\max_{p\in[0,1]} p\cdot(1-F_i(p)) \tag{3}$$

我们可以给每个影响节点提供一个单独的 OMP，从而确保仅从这一节点所得到的收益是最大的。由此提出收益最大化的第一种算法：ALL-OMP。这种算法首先计算出所有节点的 OMP，不考虑节点是否为种子节点及其影响力的大小，即对于每一个节点 $u_i\in V$，计算出 $p_i^m$ 值。然后将所有的 OMP 组成一个价格向量 $\boldsymbol{p}^{\mathrm{m}}=(p_1^m,\cdots,p_{|V|}^m)$，用 U-Greedy 算法来选择种子，当没有节点能够提供正向边际收益时算法结束。

**FFS**：通常来说，当考虑从种子节点上直接获取利润和从非种子节点上获取长远利益时，二者之间需要权衡，如果更看重后者，则产生了第二种算法：FFS。此算法向种子节点提供更大的折扣，并且向非种子节点收费。FFS 首先根据公式（3）计算 $\boldsymbol{p}^{\mathrm{m}}=(p_1^m,\cdots,p_{|V|}^m)$，之后开始 U-Greedy 算法。在每次迭代中，由卖方提供一个较大折扣（如设置价格为 0）把能够达到边际收益最大的节点加入 $S$ 集合，

当没有种子能够提供正向边际收益时迭代结束。

相比于 ALL-OMP，FFS 对种子节点的折扣持完全不同的态度。直观上来说，FFS 应该更适合高影响网络（即网络中各节点间的影响权值比较高）和低购置成本（即商品本身的价值比较低），但是对于低影响网络和高购置成本，它的表现则过于激进。例如，在图 4-4 中，当每条边的影响力权值为 0.5、购置成本 $c_a$=0.001 时，将节点 1 设为种子节点后 FFS 的利润为 0.625，优于 OMP 的利润 0.5615；如果影响力权值都为 0.01，且购置成本 $c_a$=0.01 时，OMP 的利润为 0.246，而 FFS 的利润变为 0.0025。

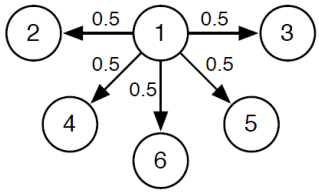



图 4-4　边影响力权值为 0.5 的节点图

**PAGE：**ALL-OMP 和 FFS 算法对于营销公司来说操作方便，但是从上述例子来看，它们的性能并不均衡也不稳定，因此可以使用 PAGE 算法。PAGE 算法依然利用 U-Greedy 框架选择种子节点，然后将所有节点的价格初始化为其 OMP 值。在每一轮迭代中，计算能够使得边际收益最大的每个候选种子节点的最佳价格，然后选择能够达到最大边际收益的节点作为种子，直到没有节点可以被选中则停止。对于所有的非种子节点，PAGE 还是将 OMP 作为它们的定价。

## 4.5　固定价格销售策略

本节收益最大化策略是向一组人（$S$）提供免费商品，对其余的人设置一个固定的价格（$p$）。然而，收入优化带来两方面的挑战。首先，种子集合 $S$ 和价格 $p$

是成对出现的，所以必须同时考虑两者，找出折中的办法：扩大 $S$ 集合会失去可能从 $S$ 中获得的潜在收益，但是可能会使得不在集合 $S$ 中的买方的正外部性增加，并可能使得卖方从中提取更多的收入。更微妙的是，在动态接受商品的过程中，对于一个固定集合 $S$ 和价格 $p$，若一个买家 $j\notin S$ 在价格为 $p$ 的情况下最初不愿意购买某商品，但是之后可能会像其他买家（不在 $S$ 中但是愿意支付价格 $p$ 购买该商品）一样支付 $p$ 的价格购买该商品，从而导致了销售的级联。

固定价格销售策略包含两个阶段。

（1）初步影响：这个阶段，卖方向一个买方集合 $A$ 提供免费商品。

（2）价格设定：这个阶段，卖方对商品设定一个固定价格。

在设定价格 $p$ 后，满足 $v_i\geqslant p$ 的买方 $i$ 会购买某商品，令 $S_1$ 为经过影响阶段后 $v_i(A)\geqslant p$ 的买方节点集合，$S_1=\{i\notin A|v_i(A)\geqslant p\}$。集合 $S_1$ 中的买方在价格为 $p$ 时购买了该商品后，他们可能会影响其他买方，从而导致其他人的估值增加并且超过 $p$，令 $S_2=\{i\notin A\cup S_1|v_i(A\cup S_1)\geqslant p\}$。随着购买者数量的增加，会有更多的买方有动力去购买该商品。这个过程不断进行并且动态变化不断传播。例如，对于任意买家 $i(2\leqslant i\leqslant k)$，给出所有已经接受该商品的买方$(\cup_{j<i}S_j)\cup A$，其中买方的估值大于或等于 $p$，$S_i$ 是不在$(\cup_{j<i}S_j)\cup A$ 中的。卖方的目标是找到一组买家 $A$ 和固定价格 $p$，从而获得最大收益。

## 4.6 商品数量受限时的收益最大化

### 4.6.1 问题陈述

本节目标为考虑商品数量有限情况下的收益最大化问题，简称 $\mathrm{RM}_{w/QC}$ 问题[18]。为了解决这个问题，这里介绍两种算法：第一种算法为策略搜索算法

（PRUB），这种算法能够得到最佳的解决方案；在 PRUB 算法之上，提出一个启发式算法 PRUB+IF，在大规模用户的情况下可以更有效地解决问题。

首先需要明确，买方的估值由商品固有价值和他人影响共同决定。图 4-5 为一个例子，每个节点代表一个用户，节点中包含的数字代表其固有价值，两个节点之间的数字代表权重（例如，节点间影响力）。为了便于说明，令 $F(x)=x$。如果图 4-5 中的用户 $f$ 购买了某商品，则用户 $c$ 的值增加到\$3+$F$(1)=\$4。如果用户 $a$、$e$ 和 $f$ 都购买了某商品，则用户 $c$ 的增长值为 $F$(3+2+1)=\$6，$c$ 的最终值为\$3+\$6=\$9。

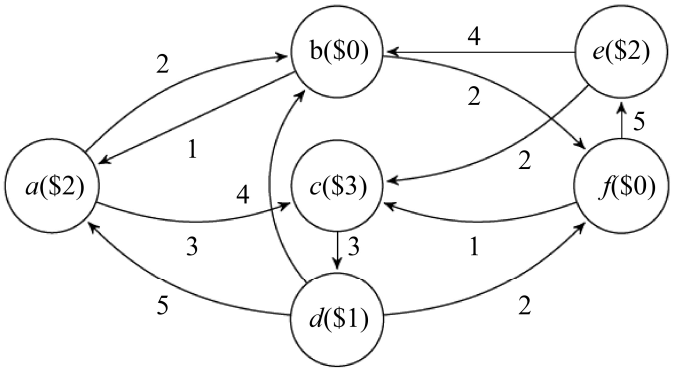



图 4-5　具有货币概念的网络

研究收益最大化问题的目标在于决定出商品的价格和找到一个可以促进销售的买家的种子集合，从而使得总收益达到最大值。

**例 1：**需要注意的是，在实际情况中，商品数量往往会受到约束，稀缺商品会更受期待[16]，例如公司会喜欢发行限量版商品来吸引买方。另一个例子是演唱会门票的促销，由于座位数量是固定值，门票的数量也会有所限制。在这种情况下，门票的数量无法随意增加，提供免费商品会减少可出售商品的数量，因此之前商品数量无限制的定价策略不适用于此场景[17]。同样以图 4-5 中的网络为例，若不考虑商品数量，在定价为\$7 时，当种子节点为{$d$,$f$}时最大收益为\$28。在{$d$,$f$}的影响下，$a$ 和 $e$ 的值都达到了\$7，成为最先购买商品的人。之后，$b$ 的值达到了

\$0+$F$(2+4+4)=\$10，$c$ 的值达到了\$3+$F$(3+2+1)=\$9，他们成为下一批购买商品的人。然而，如果商品数量被限制在 4 个，则在$\{d,f\}$的影响下只有 $a$ 和 $e$ 能够购买此商品，因此收益变为\$14。在这种情况下，当价格定为\$6 时，选择种子节点为 $d$ 能够达到最大收益。在 $d$ 的影响下，$a$ 的值达到\$2+$F$(5)=\$7，成为第一个购买商品的人；接着，$b$ 的值达到\$0+$F$(2+4)=\$6，$c$ 的值达到\$3+$F$(3)=\$6；最终 $a$、$b$、$c$ 在商品价值为\$6 时购买了该商品，商家总收益达到了\$18。因此，当商品数量受限时，需要一种新的方法来解决收益最大化问题。

## 4.6.2 PRUB 算法

要解决 $RM_{w/QC}$ 问题，Teng 等人提出了 PRUB 算法[18]。正如前文所述，问题的关键是找到一个合适的定价和种子节点集合，因为商家的收益来自商品价格和购买商品的人数。最简单的方式是枚举所有可能的价格 $p\in P$ 和所有的种子节点集合 $A$，计算出对应的收益 $R(n,p,A)$，从而找出能达到最大收益的价格 $p_{\max}$ 和种子节点集合 $A_{\max}$。然而枚举方法使得算法搜索空间太大。可以通过两种方法修剪搜索空间：（1）过滤不合适的价格（非候选价格过滤）；（2）为每个价格设定种子数量的上限（最大收益上限）。下面分别进行介绍。

（1）过滤不合适的价格（非候选价格过滤）；当商品数量限制为 $n$ 时，对于每一个价格 $p$，都有一个最大收益 $R_{\text{bound}}(n,p)$，并且有一个全局最大收益 $r_{\text{global}}$，它记录着在逐步过滤掉不能产生更高收益的价格后能得到的最大收益。若在价格为 $p$ 时的最大收益大于全局最大收益，则全局最大收益等于价格为 $p$ 时的最大收益，接着逐步过滤掉不会导致更高收益的价格。随着 $r_{\text{global}}$ 不断更新变大，不合适的价格会随之渐渐被过滤掉。

为了推断在某一特定价位时的最大收益上限，需要知道能在这个价格购买商品的买方数量；要了解一个人是否有购买潜力，需要估计出这个人的心理价

位。下面定义买家的最大心理价位，通过介绍最大值和潜在买家来定义最大收益上限。

**最大值：**买家 $v$ 的最大估价是经由 $v$ 的邻居影响后得到的估价，我们把它记为

$$X_{\max}(v)=\chi_v+F(\sum_{u\in v'\sin-\text{neighbor}} w_{uv}) \tag{4}$$

**潜在买家：**在一个特定价格 $p$ 下，买家具有购买该商品的潜力则为潜在买家。

$$X_{\max}(v)\geqslant p \tag{5}$$

**商品数量限制为 $n$、价格为 $p$ 时的最大收益上限：**

$$R_{\text{bound}}(n,p)=p\times\min\{n,m_p\} \tag{6}$$

**例 2：**以图 4-5 为例，给出凹影响函数 $F(x)=x$ 和一组价格 $p$，若销售商品为演唱会门票，假设门票总量为 4，表 4-1（a）显示了每个人的最大值，（b）则是收益最大值上限。

表 4-1 （a）最大值；（b）收益最大值上限

（a）

| $v$ | $X_{\max}(v)$ |
|---|---|
| $a$ | \$8 |
| $b$ | \$10 |
| $c$ | \$9 |
| $d$ | \$4 |
| $e$ | \$7 |
| $f$ | \$4 |

（b）

| $p$ | |
|---|---|
| \$1 | \$4 |
| \$2 | \$8 |
| \$3 | \$12 |
| \$4 | \$14 |
| \$5 | \$20 |
| \$6 | \$24 |
| \$7 | \$28 |
| \$8 | \$24 |
| \$9 | \$18 |
| \$10 | \$10 |

（2）为每个价格设定种子数量的上限（最大收益上限）：修剪搜索空间的第二个想法是，在某个价格下避免无用的种子组。PRUB 以递减的顺序来搜索每个价格的收益上限，其中收益上限小于或等于已获得的总收益的价格可以被忽略掉。在一个特定价格 $p$ 下，为了找到更大的全局最大收益，有 $\frac{r_{\text{global}}}{p}$ 个商品等待被出售，因此将种子集合数量限制如下。

**种子集和数量：**
$$|A|<n-\frac{r_{\text{global}}}{p} \tag{7}$$

按照例 2，假设初始 $r_{\text{global}}$ 为\$10。为了找到比\$10 更高的收益，首先令 $p$=\$7，PRUB 期望有多于 $\frac{10}{7}$ 张票可以卖，仅仅选择种子集合大小小于或等于 2 的集合（因为门票总数为 4），因此大小为 3 或 4 的种子集合被过滤掉。

两种修剪方法都能够得到令人满意的精确度，因此，即使搜索空间变小，算法仍然能够得到最优解。下面用一个例子讲解 PRUB 算法如何找出 $p_{\max}$ 和 $A_{\max}$。

**例 3：**首先初始化 $p_{\max=0}$、$A_{\max}$=null 和 $r_{\text{global}}$=0，根据表 4-1 中的最大收益上限，PRUB 将会按照 $p$=\$7、$p$=\$6、$p$=\$8……的顺序依次访问 $p$。从 $p$=\$7 开始，PRUB 首先检查 $R_{\text{bound}}(4,\$7)>r_{\text{global}}$ 是否成立，由于\$28>\$0，满足条件，PRUB 在价格\$7 时找到所有满足 $|A|<n-\frac{r_{\text{global}}}{p}$ 的种子节点集合（包括大小为 0 的集合），当 $|A|=0$ 时，没有买家购买商品。当集合大小为 1 时，由于 $1<4-\frac{0}{7}=4$，PRUB 列出所有大小为 1 的节点集合，此时最大收益为 $R(4,\$7,\{d\})$=\$7。PRUB 更新 $p_{\max}$=\$7，$A_{\max}=\{d\}$，$r_{\text{global}}$=\$7。由于 $2<4-\frac{7}{7}=3$，接下来寻找大小为 2 的节点组，此时最大收益为 $R(4,\$7,\{d,f\})$。PRUB 更新 $p_{\max}$=\$7，$A_{\max}=\{d,f\}$，$r_{\text{global}}$=\$14。之后，由于 $3>4-\frac{14}{7}=2$，价格为\$7 时的搜索结束。接下来考虑价格为\$6 的情况，过程与\$7 相似，最终最大收益为 $R(4,\$6,\{d\})$=\$18。$p_{\max}$=\$6、$A_{\max}=\{d\}$为最优选择。

### 4.6.3　PRUB+IF 算法

在此小节中介绍另一种可行的解决方法：PRUB+IF（带有重要反馈的 PRUB 算法），它与 PRUB 算法的不同之处在于寻找每个价格下合适的买家种子集合。继 PRUB 算法后，PRUB+IF 算法提出了价格–敏感重要性的概念，从出边邻居的反馈行为来选择种子节点，而不是列出所有的种子组。

PRUB+IF 算法的主要思想是：有更大潜力促使其他人购买商品的人更应当受到重视。贪婪地选择最重要的人作为种子节点是一个有效可行的解决方案。

问题是如何评估一个人促使他人购买该商品的潜力。一个直接的想法是统计多少人会受到此人鼓励，以及被影响的人的估价会增加多少，然后将这些信息累加起来计算此人的重要性。在这里只有潜在购买者的影响需要被累加进来，因为只有潜在购买者才具有购买该商品的可能性。此外需要注意，一个人影响另一个人购买了某种商品，另一个人会进一步影响其他人从而扩散最初那个人的影响力。为了更好地衡量一个人的重要性，级联影响也需要被考虑在内。因此，PRUB+IF 算法引入了价格–敏感重要性，包括归一化权重、影响力级联传播反馈和潜在买家过滤。

下面介绍在衡量价格–敏感重要性时的三个关键点：归一化权重、影响力级联传播反馈和潜在买家过滤。简单地说，对于每个用户 $u$：（1）归一化权重用来评价 $u$ 对其他人的直接影响；（2）影响力级联传播反馈是考虑 $v$ 受到 $u$ 影响后对其他节点的间接影响从而评价 $u$ 的潜在影响力的；（3）潜在买家过滤通过累加所有潜在买家的直接或间接影响力来提取 $u$ 的价格–敏感重要性。

## 4.7 本章小结

本章第一节介绍了最佳营销策略模型，并简单介绍了正外部性和影响-拓展营销策略；第二节对模型影响-拓展营销策略进行了简单的效率分析；第三节介绍了固定价格销售策略；最后一节介绍了商品数量受限时的收益最大化，并对 PRUB 和 PRUB+IF 算法进行了简单的描述。

相对于影响力最大化问题，收益最大化问题能够解决更加实际的问题，其精确度及效率问题值得更进一步的研究。

## 参考文献

[1] Hartline J, Mirrokni V S, Sundararajan. M. Optimal Marketing Strategies over Social Networks[C]. the 17th International World Wide Web Conference (WWW08), 2010: 189-198.

[2] D. Arthur, R. Motwani, A. Sharma, and Y. Xu. Pricing strategies for viral marketing on social networks. Proc. of the 5th Workshop on Internet and Network Economics (WINE '09), LNCS 5929, 2009: 101-112.

[3] H. Akhlaghpour, M. Ghodsi, N. Haghpanah, V.S. Mirrokni, H. Mahini, and A. Nikzad. Optimal iterative pricing over social networks. Proc. of the 6th Workshop on Internet and Network Economics (WINE '10), LNCS 6484, 2010: 415-423.

[4] N. Anari, S. Ehsani, M. Ghodsi, N. Haghpanah, N. Immorlica, H. Mahini, and V.S. Mirrokni. Equilibrium pricing with positive externalities. Proc. of the 6th

Workshop on Internet and Network Economics (WINE ’10), LNCS 6484, 2010: 424-431.

[5] W. Chen, P. Lu, X. Sun, Y. Wang, and Z.A. Zhu. Pricing in social networks: Equilibrium and revenue maximization. CoRR, abs/1007.1501, 2010.

[6] O. Candogan, K. Bimpikis, and A. Ozdaglar. Optimal pricing in the presence of local network effects. Proc. of the 6th Workshop on Internet and Network Economics (WINE ’10), LNCS 6484, 2010: 118-132.

[7] Ed Oswald. http://www.betanews.com/article/Google Buy MySpace Ads for 900m/1155050350.

[8] Katharine Q. Seeyle.

[9] Tim Weber. http://news.bbc.co.uk/1/hi/business/6305957.stm?lsf.

[10] Everett Rogers. Diffusion of Innovations, 5th Edition. Free Press, August 2003.

[11] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. Mathematical Programming, 1978, 14: 265-294.

[12] Refael Hassin and Shlomi Rubinstein. Approximations for the maximum acyclic subgraph problem. Information Processing Letters, 1994, 51(3):133-140.

[13] R. D. Kleinberg and F. T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In FOCS, 2003: 594-605.

[14] P. Domingos and M. Richardson. Mining the network value of customers. In KDD, 2001: 57-66.

[15] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In KDD, 2002: 61-70.

[16] S. Worchel, J. Lee, and A. Adewole. Effects of Supply and Demand on Ratings of Object Value. Journal of Personality and Social Psychology 32.5 (1975): 906.

[17] V. S. Mirrokni, S. Roch, and M. Sundararajan. On Fixed-Price Marketing for Goods with Positive Network Externalities. In WINE, 2012.

[18] Teng Y W, Tai C H, Yu P S, et al. An Effective Marketing Strategy for Revenue Maximization with a Quantity Constraint[C]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

# 下篇

## 工程实践

# 第 5 章　舆情监测

## 5.1　引言

舆情是人们在一定阶段和地理范围内对社会事件、社会现象产生的情感反映集合。人们对待某一事件的认知、情感、态度、行为集中地体现在社交网络舆情之中。随着互联网的发展，传统媒介对于信息的传播能力在不断地被互联网削弱，互联网以其惊人的内容创造能力和信息传播能力使得自身的重要性不断增长，传统的广播、电视、报纸、杂志因其传播能力有限、消息面狭窄而逐步让位于互联网。互联网中的信息传播是病毒式的，不断地复制、传播，使其真正实现了“一传十，十传百”的传播效果，这令传统媒介自叹不如。互联网不仅在传播上实现了超越，同时也增加了网民抒发自我情感、阐述自身态度的途径。网络是一个大的论坛，任何有表达欲望的个人都可以合理合法地实现倾诉的欲望。

虽然社交网络已经成为收集民意、了解政府和企业工作成效的有效途径，然而如果缺乏对社交网络发帖等行为的有效监管，在舆情危机事件发生后，难以及时、有效地获取深层次、高质量的网络舆情信息，经常造成舆情危机事件处置工作的被动。于是，重视对互联网舆情的应对，建立起“监测、响应、总结、归档”的舆情应对体系成了大数据时代的重要内容之一[1]。

在此背景下，舆情监测及分析行业就是为适应大数据时代的舆情监测和服务而发展起来的。其主要专注于通过海量信息采集、智能语义分析、自然语言处理、数据挖掘，以及机器学习等技术，不间断地监控网站、论坛、博客、微博、平面媒体、微信等信息，及时、全面、准确地掌握各种信息和网络动向，从浩瀚的大数据宇宙中发掘事件苗头、归纳舆论观点倾向、掌握公众态度情绪，并结合历史上的类似事件进行趋势预测和应对建议。

## 5.2　舆情监测相关技术

为了更好地理解互联网舆情监测分析技术，需要明确几个相关概念。

**（1）事件（Event）**：在特定时间、特定地点发生的事情。

**（2）主题（Topic）**：也称为话题，指一个种子主题或活动及与它直接相关的主题或活动。

**（3）专题（Subject）**：涵盖多个类似的具体事件的集合或根本不涉及任何具体事件。

**（4）热点**：热点和主题的概念比较接近，但有所区别，其主要特点如下。

① 通常是一个主题，包含种子事件及相关报道。

② 和时间相关，通常指某段时间内的热点，例如当天热点、一周内热点。

③ 和某段时间内的文档数量相关。

热点可以分为绝对热点和相对热点。绝对热点为在某段时间内文档数量超过某个固定值的主题；相对热点为按照某种排序方式排名靠前的若干个主题。

**（5）主题监测**：从信息流中自动监测出最新的主题，并将报道及时按照主题组织起来。

**（6）热点自动发现**：也叫作热点监测，就是如何从不断涌现的网上舆情中及时发现新发生的热点信息，并对其进行持续追踪。热点监测任务可以在主题监测任务的基础之上加入时间和数量两个因素的分析来解决热点发现的问题。

**（7）热点分析**：在热点自动发现任务的基础上，对自动发现的热点进行深入分析，从多方面、多角度综合分析和展现当前的舆情热点。研究内容包括舆情热点的关键词和摘要提取、情感分析、传播分析、趋势分析、关联分析等任务。

（8）**报道**：指一个与主题紧密相关的、包含多条陈述某事件的句子的新闻片段。

互联网舆情监测分析的主要目标是在主题发现和追踪技术的基础上，通过自动发现和深入分析的方式综合展现当前的舆情热点，其主要研究内容包括以下几个方面。

**（1）舆情热点自动发现**：就是基于主题监测技术帮助人们应对信息过载问题的研究，以互联网新闻、论坛、博客等媒体网页作为处理对象，自动发现对新主题的报道，并将涉及某个主题的报道组织起来以某种方式呈现给用户。其目标是实现按主题查找、组织和利用来自多种信息源的信息。本技术可以提高舆情监测的综合性，实现对多种来源、多种形式的舆情的综合性分析和监测能力，为全面掌握新闻、论坛、博客等各种网络传播媒介的舆情热点、传播动向、趋势分析等提供基础。

**（2）舆情热点的关键词和摘要提取**：就是自动从热点的文档集合中摘取精要

或要点，其目的是通过对原文本进行压缩、提炼，为用户提供简明扼要的内容描述。而关键词和摘要都是描述一篇文章或一个文档集主要内容的重要部分，不同之处在于摘要中提供的是语义连贯的句子，而关键词抽取的是彼此独立的词汇。本技术可以为文档或文档集生成高质量的关键词和摘要，方便用户浏览检索结果或文档集合，了解文档或文档集内容。

**（3）舆情热点的褒贬分析**：就是对热点内的文档和回复信息进行褒贬分析，通过分析褒义词和贬义词，结合上下文进行语境分析，或者通过基于机器学习的褒贬分析算法计算出文档和回复的褒贬因素。在得到褒贬因素的同时，可以加权给出每篇文档的褒贬因素度量值，再按时间统计出该热点的总褒贬指标变化及某一段时间范围内的褒贬指标增量。当褒贬指标超出某一安全范围时可以给出提示信息，用于舆情信息的提前预警。

**（4）舆情热点的传播动态分析**：就是利用博客、论坛、新闻等关联分析技术，实现对某个主题传播趋势的分析，用动态传播图的形式展现舆情传播的线索。舆情传播动态模块对同一主题的论坛帖文、博客文章、网站新闻进行基于时间的罚分策略从而进行关联程度分析，以传播网的形式给出同一主题在不同媒介之间的传播关系，结合关注程度分析得出主题的转移趋势，并以平面图、传播动画及抽象的有向图传播示意图展现给用户。

**（5）舆情热点的趋势分析和关联分析**：通过三维图形下的信息挖掘、叠加检索模型，以及概念挖掘手段，以波谱图的方式，展现一定时间周期内的舆情变化情况，以及舆情重点和相关关系。系统通过粗细、亮暗、分叉的方式来表达同一时期报道信息的数量、关注度、趋势等，为舆情变化判断提供一定的参考。

## 5.2.1　舆情热点自动监测设计

舆情热点监测技术就是从网络上不断涌现的舆情中及时获得新发生的热点事

件信息，并对其进行持续追踪。主题监测与追踪技术是解决这一问题的基础，本章要解决的问题是改进现有的主题监测方法用于热点监测。

### 1．主题监测

主题监测就是从新闻信息流中自动监测出各个主题，将每篇新闻报道划归到相应的主题，并且能够实时地针对新到的新闻报道监测新的主题。

主题监测算法是对文本聚类算法的改进和延伸。监测的目的就是要按照新闻报道表达的主题将其进行聚类。一般可将主题监测技术分为回溯监测和在线监测两类，它们之间有一定的差别。回溯监测的目的是从已有的新闻报道集合中发现以前未标识的新闻主题，要求系统输出新闻主题的信息，能够说明新闻报道和主题的关联关系。而在线监测的重点在于及时地从实时新闻报道流中标识新的主题，也就是在某个表达新主题的报道出现的时刻标识出该新闻主题。

主题监测技术包括新闻文本特征的选择、相似度的计算方法及核心监测算法三个方面的内容，从其中任何一个方面进行改进都有可能改进主题监测的结果。主要的主题监测算法有基于平均分组的层次聚类法和在线增量式聚类算法，其中在线增量式聚类算法能够及时地从新闻信息流中监测到新主题，应用最为广泛。

### 2．舆情热点自动监测设计

由于现有主题监测技术主要考虑在固定小数据集合上的错检率和漏检率，在实际应用于舆情热点的自动监测时，存在主题排序、主题相似性、报道淘汰和主题描述等缺陷。针对这些问题，本章介绍一种新的舆情热点监测方法，该方法利用舆情热点本身的特点，通过引入主题排序、主题合并与调整、报道淘汰及主题描述等步骤，实现对持续新闻流进行动态、高效的热点监测。

舆情热点自动监测方法具体包括以下几个步骤。

（1）从数据源读入一篇报道，对多个网络新闻数据源进行不间断的监测，从网络中自动抓取新闻报道，解析出新闻报道的时间、标题和正文信息等。如果没有从报道中找到时间，则以抓取时间为准。

由于多个数据源之间存在多次重复，对新抓取的新闻报道，根据报道的文本内容进行消重处理。如果新报道和之前已经处理的新闻报道的重复度大于某值，则认为是重复的新闻报道。

由于新闻报道的范围过于宽泛，采用基于来源的规则分类及基于内容的自动分类相结合的方法对新闻报道进行分类（类别是预先设定好的，如可以分成美国、日本、欧洲、维稳等）。规则分类根据新闻来源及作者等进行分类。基于内容的自动分类可以采用向量空间模型（VSM）和支持向量机算法（SVM），根据报道内容和标题对新闻报道进行自动分类，并且按照所属类别处理步骤（2）到步骤（7）。

（2）采用质心比较策略，将报道与所属类别内现有的新闻主题进行比较，同时考虑时间特征和内容特征，计算报道和主题间的相似度，并记录最大相似度及相似度最大的主题，确定与当前报道最相近的主题。

（3）根据步骤（2）计算得到最大相似度 $S_{\max}$ 及相似度最大的主题，对当前报道采取如下措施：

A．如果 $S_{\max}$ 小于创新阈值，则在该报道所属类别内创建一个新主题。

B．如果 $S_{\max}$ 大于创新阈值而小于聚类阈值则不作处理，返回步骤（1）。

C．如果 $S_{\max}$ 大于聚类阈值而小于贡献阈值，则归入当前主题。

D．如果 $S_{\max}$ 大于贡献阈值则归入主题，并调整上述 $S_{\max}$ 和各阈值的取值范围均大于 0 而小于等于 1。

（4）对一个类别内的新闻主题两两比较，如果两个主题的相似度大于合并阈值则将其合并。主题之间的相似度计算公式可以采用传统聚类算法中计算两个聚类相似度的方法，例如基于向量空间模型，综合考虑两个主题中所有新闻报道之间的两两相似度，采用如下公式：

$$\mathrm{Sim}(E_1,E_2)=\frac{\sum_{d_i\in E_1}\sum_{d_j\in E_2}\mathrm{Sim}(d_i,d_j)}{|E_1|\cdot|E_2|} \tag{1}$$

其中，$E_1$,$E_2$ 是两个监测到的新闻主题，$d_i$,$d_j$ 分别为 $E_1$,$E_2$ 中的新闻报道，$\mathrm{Sim}(E_1,E_2)$是两个新闻报道之间的相似度，$|E_1|\cdot|E_2|$分别为两个主题中包含的新闻报道数目的乘积。

（5）用户对各主题内的新闻报道进行淘汰，并重新计算新闻报道和该主题的相似度。对相似度低于聚类阈值或者不满足限制条件的新闻报道进行淘汰，然后再重新计算。

（6）若当前类别内的主题数量超过主题窗口大小，则对类别内的所有新闻主题进行排序。结合新闻主题的时间特性和数量特性，计算新闻主题的得分值进行排序，并且同时考虑多个不同的排序，只有当主题在任何排序中都不在主题窗口内时，才将该主题淘汰，这样多重排序就给用户提供了不同粒度的信息参考。系统将不在主题窗口内的新闻主题淘汰，有利于提高系统处理的效率。

（7）根据用户要求，对外输出监测结果。结合主题的时间特性和主题内的新闻报道数量特性，从所有类别中选出得分最高的若干个新闻主题，作为该类别最热点的新闻主题，输出主题描述和包含的新闻报道列表。其中，主题描述的生成过程如下：

A．读取主题内部权重最高的若干个特征词。

B．在与主题相似度大于主题阈值的新闻报道中，选择时间最近的一篇新闻报道的标题。

C．综合 A 和 B，输出该主题的描述。

舆情热点自动监测流程如图 5-1 所示。

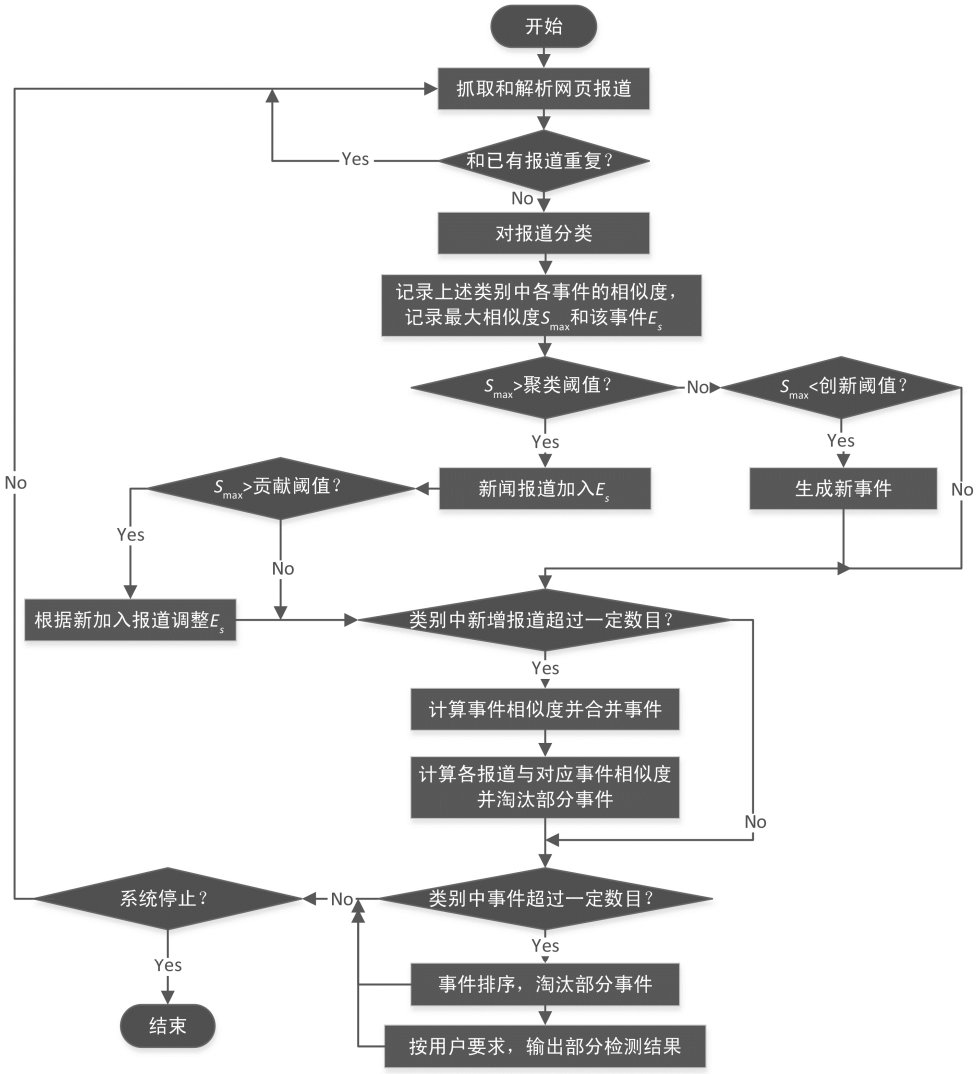



图 5-1　舆情热点自动监测流程

## 5.2.2 文档关键词提取设计

### 1. 文档关键词提取

文档关键词提取就是从文档中自动选取重要的词汇或短语。其目的是通过对原文本进行压缩、提炼，为用户提供简明扼要的内容描述。

根据不同的划分标准，文档关键词提取可以分为以下几种类型。

（1）根据处理文档的维度，关键词提取可以分为单文档关键词提取和多文档关键词提取。单文档关键词提取只对单篇文档提取关键词，而多文档关键词提取则对一个文档集提取关键词。

（2）根据所采用的方法，关键词提取可以分为生成式和抽取式。生成式方法通常需要利用自然语言理解技术对文本进行语法、语义分析，对信息进行融合，利用自然语言生成技术生成新的关键词。而抽取式方法则相对比较简单，通常利用不同方法对文本的结构单元字、词、词组等进行评价，对每个结构单元赋予一定权重，然后选择最重要的结构单元组成关键词。抽取式方法应用较为广泛，其通常采用的结构单元为单词。

（3）结合词性标注和统计信息的关键词提取方法有效地利用了多个文档所反映的全局性重要信息，同时又尽可能地过滤掉信息冗余。

### 2. 文档关键词提取技术设计

多文档关键词提取的困难之处在于需要分析多个文档所反映的全局性重要信息。此外，文档集中不同文档所包含的信息不可避免地相互重叠，因此需要有效的方法来对不同文档信息进行融合。如果有可能的话，还要比较这些文档信息的差异。也就是说，多文档关键词提取既要尽可能保留文档集中的重要信息，又要尽可能过滤信息冗余。

算法可以分解为三个步骤，即找出候选关键词、过滤并计算候选关键词权重、

给出最终关键词。首先基于词性标注结果找出名词、动词、名词短语作为候选关键词；其次过滤部分候选关键词，并计算其他候选关键词的权重；最后根据用户所需给出最终关键词列表。下面对这三个步骤进行详细介绍。

（1）找出候选关键词。

在这一阶段中还分为两个步骤：输入预处理和识别候选关键词。对文档进行切分和词性标注，并利用文章中出现的标点符号，将文本输入流分成语言段。原始的边界包括标点符号等，终极词汇包括感叹语、数字等。分词的过程中可以采用命名实体识别功能来增加分词的准确率。经过以上处理，我们采用名词、动词、名词短语作为本文的候选关键词。

实现过程中，采用有限状态自动机来识别名词短语。同时，考虑到很多时间性名词意义不太明确，所以过滤掉时间性名词，例如“当前”、“现在”、“年”、“月”等。

（2）过滤并计算候选关键词权重。

这个阶段是整个提取过程的核心，也是各种算法创新的所在，更是最能体现算法效率和精准度的关键点。

算法的过滤过程如下：对主题内出现的时间名词、单字词、长度大于 7 个汉字的词或短语、只在主题内一个文档中出现（如果主题包含两个或以上文档）的候选短语进行过滤，然后将候选关键词按照出现次数排序。对于文档内出现次数相同的候选短语，按照包含规则进行过滤，即如果短语“A”和短语“AB”出现次数相同，则过滤掉短语“A”。对于有类似包含关系但出现次数不完全相同的短语的过滤，在第三阶段给出。

计算权重阶段又可以分为选择特征和根据特征计算权重。可供选择的文档特征包括如下几点。

- 主题内的短语频率（TF）：短语在主题内总共出现次数，即短语在主题内文档中出现次数的总和。

- 主题内短语的文档频率（DF）：短语在主题内多少个文档中出现。
- 短语在整个文档集倒排文档频率（IDF）。
- 词汇在文档内第一次出现的位置和文档长度。
- 主题内词汇平均第一次出现的位置。

本文提出的算法试验了 TF 和 TF×IDF 两种计算方法。考虑到特征表达方式并不能很好地反映文档的特征，因此人们往往采用一种更好的 TF×IDF 向量表示法。这种表示法充分考虑了字词在文档集中的分布情况，能够有效地反映文档的特征。其中，IDF 计算公式如下：

$$\mathrm{IDF}(W)=\log(N/\mathrm{DF}(W)) \tag{2}$$

（3）给出最终关键词。

得到词语权重并进行排序，最终给出用户需求个数的关键词。在给出最终关键词之前，还需要根据包含规则对关键词进行二次过滤。例如按照第二阶段过滤和排序之后，“物权法”主题选出来的部分词汇列表及其词频如下：物权法/129，草案/106，财产/87，法律/84，全国/73，物权/50，人大/45，物权法草案/43，全国人大/43。

从上面的例子中可以看到，“物权法”、“物权”、“草案”三个词汇及“物权法草案”均在候选集合中，而“物权法草案”明显包含了更丰富的含义和内容，也是我们希望抽取出来的关键词。“全国”、“人大”、“全国人大”也是同样的例子。因此，在这个阶段我们的任务就是根据需要的关键词个数，去掉那些被包含的词汇，而尽量选取更长一些、含义更丰富和明确的短语。

### 5.2.3 专题生成技术分析设计

专题这个概念比主题具有更高的层次。一般认为，相互关联的不同新闻主题组成一个专题，比如“习近平访问美国”和“布什访问中国”这两个不同的新闻

主题都属于“中美关系”这个专题。性质相同的两个不同新闻主题也可以组成一个专题，比如“火山爆发”专题可以包括“维苏威火山爆发”和“夏威夷火山爆发”的相关主题。由相互关联的新闻主题组成的专题的时间特征并不是很明显，组成专题的各新闻主题之间不具有明显的时间近邻性。

因此，在未知专题信息的前提下，专题生成可以通过对监测出的新闻主题进行聚类来实现。当然，专题生成也可以直接对新闻文本进行聚类来实现，这也是目前在主题监测和追踪（TDT）中常采用的方法。这种方法的优点是简单；缺点是对新闻信息的组织层次不清晰。如果已知专题信息，需要将主题划分到已知的专题中，则可以采用分类技术来实现。

### 5.2.4　主题生成技术分析设计

#### 1．主题追踪

主题追踪是从新闻报道流中追踪那些讨论与目标新闻主题相关的报道。主题追踪技术能对特定新闻主题单独进行追踪，追踪过程中并没有关于主题的大量信息可以利用，唯一可以分析利用的只有主题训练集中的少量相关新闻报道，可能还有一些不相关的新闻报道。由于主题具有动态性，因此主题追踪是一个动态学习过程，而且要求对新闻报道流进行实时追踪，不能有延迟。

对于每一个被追踪的主题，系统实时对每篇被处理的新闻报道输出一个决策值——是或否，表明该新闻报道是否为此主题的相关报道。除了这个基本输出外，为了对追踪结果进行评测分析，系统还要对每篇新闻报道输出一个得分值，表示该新闻报道属于此主题的可信度。

最简单的主题追踪方法就是基于信息检索技术的构建查询方法。其核心思想是根据训练集中的新闻报道构建一个用来追踪的查询表达式，然后将待处理的新闻报道与该查询进行匹配。这种方法一般基于向量空间模型（VSM）。

基于分类技术的改进算法也是主要的主题追踪算法，主要分为 KNN 算法和判定树算法两种。由于大多数主题在不断演化，并且在一定时间内消失，所以要求主题追踪系统能够适应主题的动态变化，并且在适当的时间条件下停止追踪。

### 2．主题追踪的技术设计

主题追踪技术为最基本的构造查询方法，从训练集中构造一个用查询向量表示的追踪器，然后使用此追踪器在线对新闻报道做出追踪判断。

系统将自动地从一个或多个相关报道及若干个不相关报道中创建追踪器。构造追踪器的过程主要考虑三个问题：特征选择、权重赋值及闭值估计。选择的特征及其权重将组成查询向量，从而构造出追踪器。

对于新闻报道的表示仍采用向量空间模型，特征采用分词系统输出的词语，其权值由 TF 和 IDF 的乘积得到。

考虑到主题具有一定的生命周期，通常认为如果当前新闻报道与主题最后一篇相关报道之间间隔很多其他新闻报道，表明二者的时间差距过大，那么该新闻报道属于此主题的可能性很小，可以对当前新闻报道做出“否”的判断，并停止追踪。

## 5.3　互联网舆情监测分析应用系统

互联网空间每天都在产生着海量信息，网络空间已经成为民众抒发民意、组织各种活动的重要场所，也是了解舆论、监测网络活动的重要场所。但是，面对迅速增长的互联网信息，人工方式已经远不能实现对互联网信息处理和互联网舆情监测分析的现实需要。

互联网舆情监测分析系统指的是整合互联网搜索技术及信息智能处理技术，

通过对互联网海量信息自动抓取、自动分类聚类、热点发现和分析、专题聚焦等，实现对网络舆情监测和新闻专题追踪等需求，形成简报、报告、图表等分析结果，为政府部门及企业全面掌握网络舆情、争取处置主动权提供有效的分析依据。

本节主要介绍一个实用化的互联网舆情监测分析应用系统。

## 5.3.1　互联网舆情监测分析系统结构

互联网舆情监测分析系统结构如图 5-2 所示。

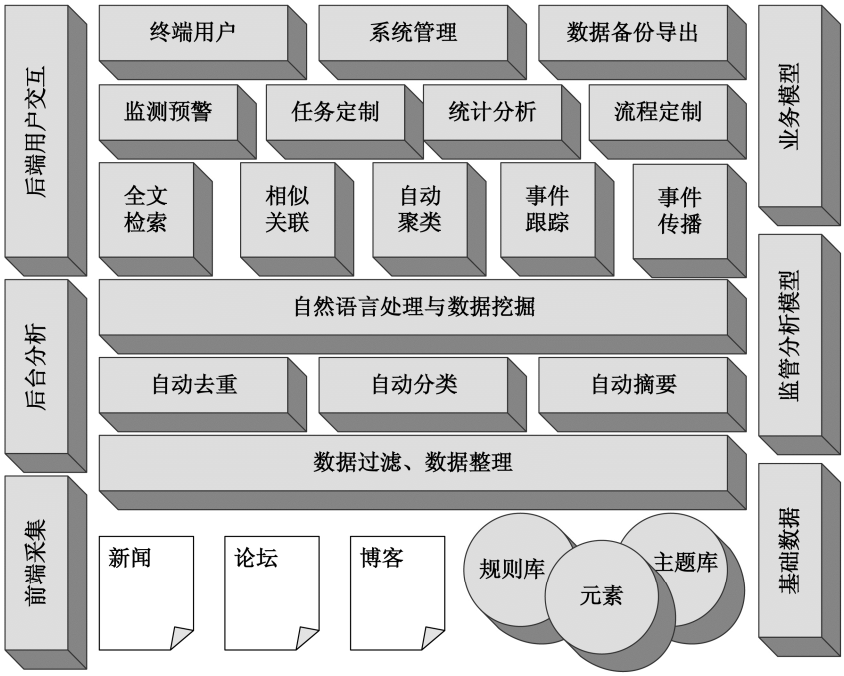



图 5-2　互联网舆情监测分析系统结构

互联网舆情监测分析系统是以中文信息处理技术与数据挖掘技术为核心技术、以智能分析和智能处理为核心功能的应用系统。其主要模块包括以下几个。

### 1．自动分类

自动分类采用支持向量机作为分类引擎，具有多层次分类功能，具备增量学习与反馈学习的能力。其主要用于对互联网新闻、论坛、博客信息的自动分类。

### 2．自动去重

自动文本去重是利用文档的内在特征信息进行智能分析，判断文档的相似性与重复性。自动去重引擎采用了文章相似度量技术与相似索引技术，适合于海量文档场合的快速相似判断。在实际应用系统中，在该引擎核心上，一方面可实现对文本的自动去重，降低文章冗余度，避免文章重复发布等；另一方面可实现自动查找相似文本并向检索者进行推荐。

### 3．自动摘要

包括单文档摘要与多文档摘要。单文档摘要方法是综合考虑句子的词频、位置等特征对句子进行权重计算，抽取重要的句子形成摘要。多文档摘要采用基于句子关系的摘要方法。多文档摘要主要用于为文档聚类及主题监测得到的主题类簇提供简洁的摘要，方便用户了解主题类簇的内容。

### 4．自动聚类

自动聚类引擎可以实现对检索结果自动聚类并构建树状结构，以使用户快速定位所需信息并对新闻稿件自动聚类，实现辅助专题制作等。系统的自动聚类引擎采用文档向量空间模型与改进的聚类算法，在处理海量文档及媒体数据时具有更好的性能。

### 5．相似关联

相似关联引擎基于倒排索引实现快速初始文档相似搜索，并利用基于文档结构的相似搜索模型对初始结果进行重排，计算文档间的相似度，指出多文档之间的内容关联关系。

### 6．事件跟踪

事件跟踪引擎采用构造查询的方法，主要根据训练集中的新闻报道构建一个用来追踪的查询表达式，然后将待处理的新闻报道与该查询进行匹配。这种方法一般基于向量空间模型。

### 7．监测预警

监测预警引擎利用舆情热点自动监测方法，可对互联网新闻、论坛、博客信息进行分析，快速在线监测网络舆情的新主题和新热点。通过系统设定的关注规则，当关注信息超过系统阈值时，系统会通过界面、声音、邮件等方式对监控信息进行实时报警。

## 5.3.2　互联网舆情监测分析系统功能

互联网舆情监测分析系统是一个监控互联网网页内容的应用系统。其任务是：高频度采集网页内容，实时监控用户关注的新闻网站、论坛、博客，根据用户关注的方向及时预警，为用户监控互联网提供手段。其主要功能包括以下几个方面。

### 1．网络数据采集

网络数据采集主要利用主动抓取工具实现通过代理和非代理模式对互联网网页进行高频度采集。抓取工具可以精确提取新闻、论坛、博客的页面元素和元素相应位置，并实现每几分钟定期轮询一次。

### 2．网络数据分析处理

网络数据分析处理是利用全文检索和数据挖掘技术，实现互联网信息的自动分类、聚类、去重。利用中文自然语言处理技术实现监控信息的关联分析、事件跟踪、传播关系等功能。系统实现了单一条件、多条件复合、同音、同义等多种检索方式。

### 3．互联网舆情热点自动监测

对互联网舆情进行精确分析，快速识别新主题和新热点。利用本文改进的舆情热点监测方法，通过引入主题排序、主题合并与调整、报道淘汰及主题描述等步骤，实现对持续新闻、论坛、博客等信息进行动态、高效的热点在线监测。当关注信息超过阈值时，系统将进行实时报警。

### 4．事件跟踪和趋势分析

从互联网数据中追踪那些讨论目标新闻主题的相关报道、论坛、博客等信息。通过主题监测与追踪技术，对主题之间相互演化、发展趋势进行分析，从海量的新闻信息中快速、准确地找到感兴趣的深层次新闻信息，为用户提供更高层次的服务。

### 5．专题自动生成

从大量的互联网新闻报道中，对主题信息进行动态聚类，形成专题信息，实现辅助专题制作等。

### 6．统计分析和系统管理

系统实现后台统计分析管理，包括统计报警数量、采集源数量、点击率、回帖数及评论数等，实现用户自定义多角度查询。系统实现日志管理、用户管理、权限管理、任务管理和数据备份等后台管理。

## 5.4　典型舆情监测系统

下面介绍由北京邮电大学可信分布式计算与服务教育部重点实验室所开发的网络智能舆情监测系统（Web-based Intelligent Public Opinion Monitoring System，IPOM）。其目标旨在建设面向全国各大热门论坛 BBS 的舆情管理工程。重点面向

各大知名高校论坛 BBS、热门网络社区等，通过有效地采集网络数据获取舆情信息，对其中的数据（包含文本、图片、视频等多种媒介）进行筛选、清洗、聚类和存储，以期达到分析、预警、推送的舆情智能处理，提供全面、及时的舆情监管、控制、预防、警报服务，建立实用、高效、有力的舆情监管工程。其建设目标如图 5-3 所示。

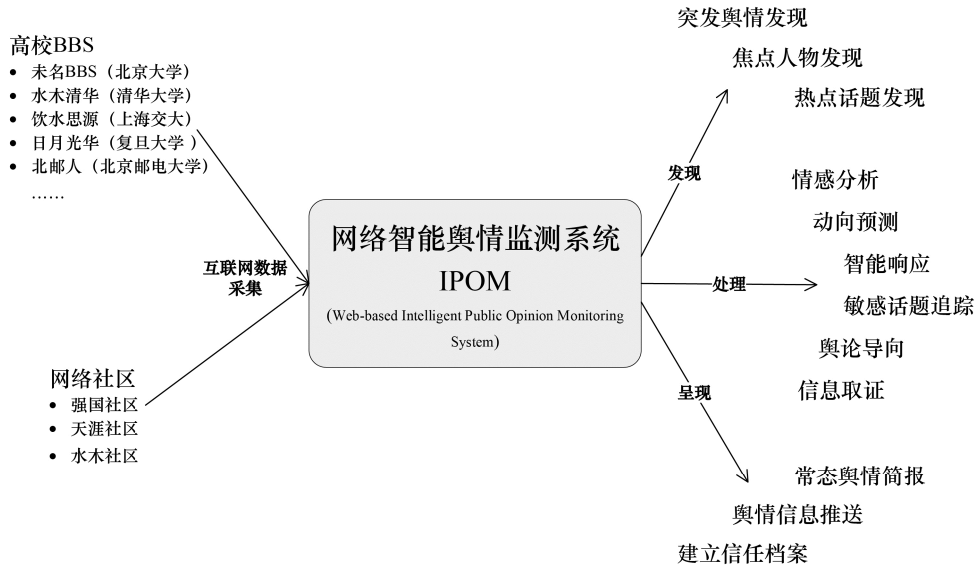



图 5-3　IPOM 建设目标图

网络智能舆情监测系统（IPOM）针对舆情这一特殊对象，业务层设计包括信息采集子系统、舆情分析子系统、舆情处理子系统及舆情呈现子系统，4 个子系统在统一管理平台的管控下合理分工、协调工作，实现舆情信息从源头采集到可信监控的全过程。另外，设计安全保障子系统保障全系统的安全性，防止黑客入侵和恶意破坏。IPOM 体系架构设计图如图 5-4 所示。

为了实现系统建设目标中的各项指标，将功能细化至 4 层业务子系统中，如图 5-5 所示。

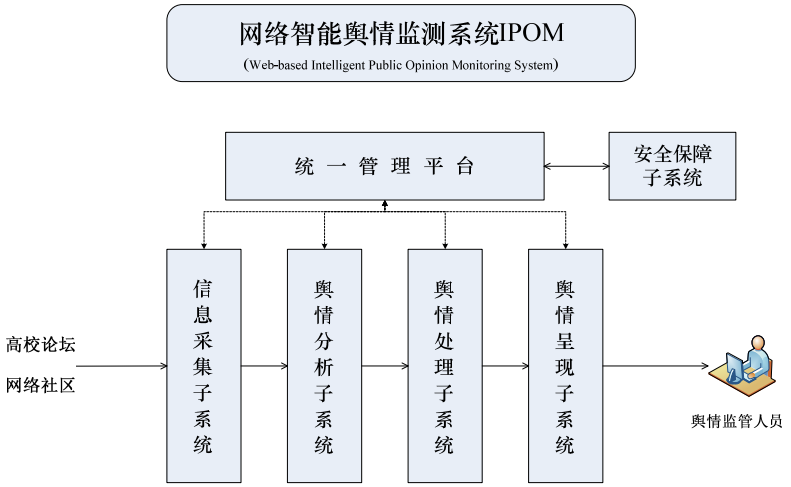



图 5-4　IPOM 体系架构设计图

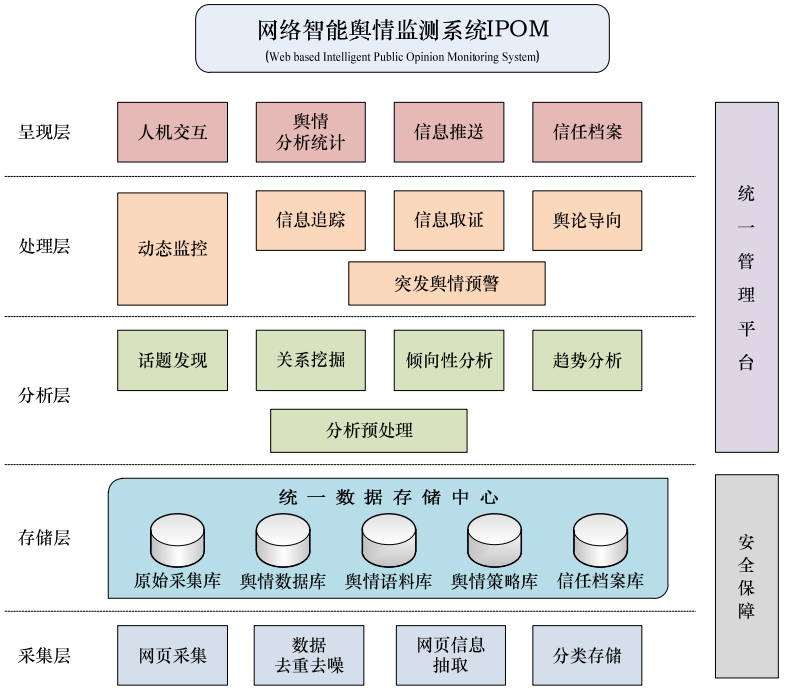



图 5-5　IPOM 自底向上层次设计图

下面针对每个系统体系架构中的每个部分逐一说明。

### 5.4.1　信息采集子系统

针对论坛信息源的数据规模大、信息更新快、链接层次深、噪声干扰大、需要特定访问权限等特点，需采用高效的信息采集方法，周期性快速获取噪声低、重复内容少并且格式统一的网页信息，建立统一、完善的舆情获取平台。信息采集子系统架构图如图 5-6 所示。

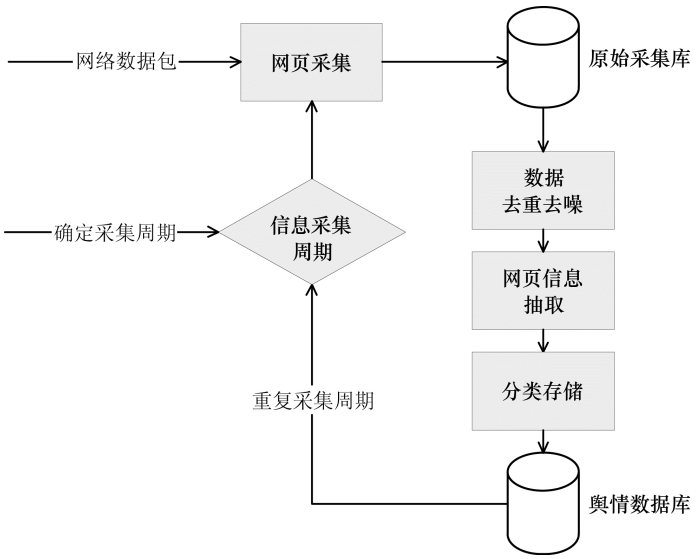



图 5-6　信息采集子系统——系统架构

#### 1．网页采集

论坛（BBS）作为网络人际交互中最活跃的数据源，具有数据规模大、信息更新快、链接层次深等特点。为了能够快速、精准地采集有用信息，有必要减少无用、干扰的网页信息。因此，在进行网页采集时，针对不同的论坛，可以选取不同的采集途径；由于动静态页面的实现技术不同，信息采集的方法也会不同；为了保证所采集信息的完整性和实时性，需要采用增量采集的方法，

配合不同论坛的信息采集周期进行信息抓取，同时对帖子的原貌进行快照处理并保存。

### 2．数据去重去噪

（1）数据去噪：通过页面去噪，把网页中与舆论分析不相关的内容识别并去除，只保留可供挖掘和参考的信息。例如，在建立全文索引的预处理环节需要把广告和相关链接部分识别为噪声并去除，以便减少索引部分的容量和提高搜索的质量等。

（2）数据去重：众多高校的学生之间会互相访问论坛，跨论坛发帖和回帖会导致信息采集时采集到重复的 URL。这会造成对同一帖子的重复采集，将会浪费大量的存储空间并影响论坛数据的处理效率，并且冗余信息会对舆情数据造成干扰，进而影响到最终舆情分析的效果。所以，在进一步对论坛上的网页信息进行数据处理之前，需要自动对网页进行去重处理。

### 3．网页信息抽取

由于网络的开放性、动态性与异构性等特点，通过定向搜索网络论坛得到的网页数据是半结构化的、分散的、异构的、没有统一管理的，而且布局风格和内容变化迅速。要想提高后续舆情分析等一系列复杂工作的质量与效果，必须首先对定向搜索得到的网页数据进行信息抽取，从半结构化文档中抽取出结构化信息，并将数据存入数据库中以便后续的分析、处理。

### 4．分类存储

由于数据量庞大，为了便于后续的工作中对数据的按需查询，此系统对海量数据进行了分类存储。分类后的数据不仅简化了数据的检索，而且有利于后续的舆情分析等工作。

## 5.4.2　舆情分析子系统

舆情分析子系统的主要目标是对全国各大高校 BBS 及热门网络社区采集的数据进行分析预处理，初期主要针对文本数据，及时发现焦点人物及敏感、热点话题，挖掘话题包含的潜在关系，并对话题及其相关评论进行倾向分析和趋势分析，同时将热门话题涉及的人物、事件、地点、时间等信息更新至舆情语料库。舆情分析子系统架构图如图 5-7 所示。

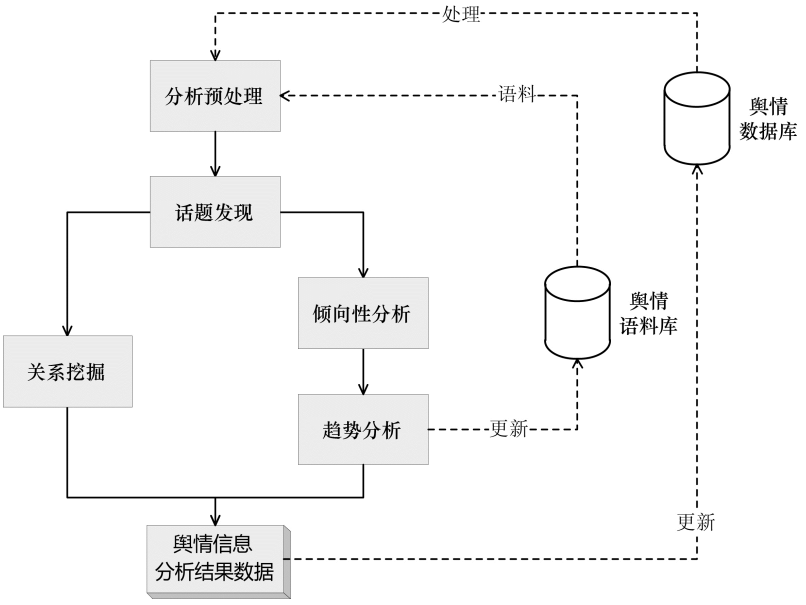



图 5-7　舆情分析子系统——系统架构

### 1．分析预处理

分析预处理主要是对全国各大高校 BBS 及热门网络社区采集的文本数据进行基于语义分词和句法分析。语义分词主要是将文本转换成词条，这些词条是未来话题特征的主要来源，所以说分词是话题发现的基础，便于后期识别敏感词及话题关键字。分词主要包括以下几个任务：词典查询、词语切分、未登录词识别、

切分歧义排除。句法分析则旨在识别时间的具体信息，提取相关倾向性词语及倾向对象。

### 2．话题发现

为了实现对全国各大高校 BBS 及热门网络社区中帖子动态的有效监控，需要进行话题发现。话题发现本质上就是将一个文本集分组的自动处理过程，采用文本聚类的方法进行自动分组。舆情分析子系统的话题发现主要包括敏感话题发现和热点话题发现。

### 3．关系挖掘

舆情追踪和取证有时需要以舆情潜在关系为线索。话题包含的潜在关系主要有三种：人物与人物之间的关系、事件与事件之间的关系、人物与事件之间的关系。通过对话题相关主帖和回帖内容进行分析，挖掘主要潜在关系，为舆情追踪和取证提供依据。关系挖掘是一种特殊形式的实体关系抽取，能够从纯文本中发现实体对之间存在的语义关系。主要操作有实体识别、关系提取、单文本指代消解、跨文本指代消解及关系融合。

### 4．倾向性分析

网络信息和社会信息的交融对社会的直接影响越来越大，但由于网络上的信息量十分庞大，仅靠人工难以应对网上海量信息的收集和处理，因此需要依靠倾向性分析技术自动地对舆情信息进行监控。文本倾向性分析过程主要针对带有情感色彩的主观性文本进行分析、处理、归纳和推理。倾向性分析可以为帖子设置情感标签，了解和归纳用户的主流观点，并进行分析和统计，为话题动向分析和舆情处理决策提供支持。主要操作有倾向性词语的抽取和判别、倾向对象的抽取。

### 5．趋势分析

每个话题都有一个产生、发展、结束的过程，舆情监控系统倾向于跟踪话题

的来龙去脉，了解该话题的全貌。在一个话题发展的每个时间点都会有一些相关的用户发帖，而这些就相当于对该话题进一步的发展和描述。话题趋势分析就是通过对话题相关主帖及回帖的分析来对话题的发展进行追踪。话题的描述方式为二维热度图，横坐标表示时间点，纵坐标表示某时间点上话题讨论的数量，其中涉及对正、负向观点的百分比统计。

#### 6．舆情语料库

舆情语料库主要包括两部分内容：一是敏感词库，里面记录的是各类敏感话题涉及的特征词或关键字；二是热门词库，存储了各类热门话题涉及的特征词或关键字。话题发现会调用舆情语料库中的数据，快速识别敏感话题和某一时间段内的热门话题。根据趋势分析得到的各话题的热度走势及倾向性分布情况，将新发现的热门话题涉及的人物、事件、地点、时间等信息更新至舆情语料库，便于对相关话题进行有效识别和监控。

### 5.4.3　舆情处理子系统

舆情处理子系统的设计主要包括动态监控模块、信息追踪与取证模块、突发预警模块和舆论导向模块。动态监控模块负责舆情话题的实时监测，掌握话题趋势的发展与变化情况，根据舆情策略库中存储的各种处理策略，及时做出相应的智能处理，从而对网络舆情的发展态势进行有效的控制。其中，动态监控的智能处理包括追踪敏感话题的发展路径，并取证为后续工作提供评判的依据，及时进行舆情引导，当舆情的发展态势评测结果达到预设的预警阈值时，系统自动向系统管理员及高层监管人员发出突发舆情预警提示，相关管理人员收到预警后，及时做出相应的应对处理。舆情处理子系统架构图如图 5-8 所示。

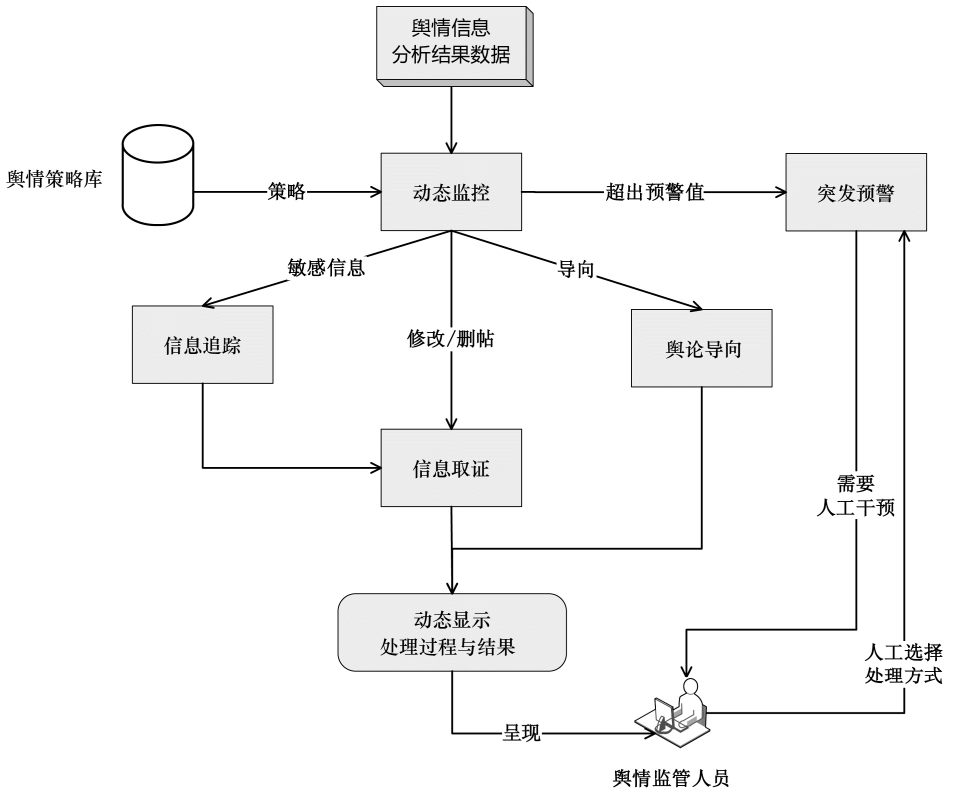



图 5-8　舆情处理子系统——系统架构

### 1．动态监控

基于舆情分析层处理得到的数据，针对敏感专题和特定舆情事件进行实时监控和持续跟踪，随时掌握其发展和变化趋势，及时发现在各个信息通道中发布、传播的敏感信息。一旦监控结果达到预警的条件，动态监控模块将自动调用相关的舆情处理模块对网络舆情事态的发展进行实时、有效的监控。其主要功能包括：实时监控、动向预测、持续跟踪。

### 2．舆情策略库

舆情处理层对舆情的生成、发展和变化进行有效的监测，并实时分析判断舆情的发展趋势，通过智能决策及时做出智能处理，动态监控网络舆情的发展态势。

舆情策略库负责存储需要发送给动态监控的各种策略信息，帮助动态监控进行智能决策，为舆情处理层的智能决策提供依据。

### 3．突发预警

突发预警基于话题分析的结果及话题趋势预测的结果，及早发现可能产生重大影响的舆情话题，并及时对这些舆情话题的走向、发展规模进行判断，一旦舆情话题的发展态势达到预先设定的预警条件，系统将自动做出预警，从而在话题大规模传播造成恶性影响之前采取相应的应对措施，对话题的发展进行有效的调控，避免可能对国家安全、社会稳定造成的影响。

### 4．信息追踪

通过信息追踪对敏感热点话题进行跟踪和追溯，可以找到舆情事件的源头，了解网络不良信息的扩散趋势和整个扩散过程，掌握舆情事件的发展全貌，也为舆情的智能决策提供可靠、全面的依据。主要功能包括：对关注度、影响力等指标的综合分析，实时关注话题内容的动态变化及其派生内容，实时追踪话题发展路径和人物溯源等。

### 5．信息取证

信息取证是对敏感不良舆情信息进行收集、提取证据的过程。通过对敏感热点话题进行取证来收集相关的舆情证据，从而为舆论的决策和进一步监管提供依据，同时也为有关非法行为的评判提供法律依据。主要包括对敏感和不良信息的电子取证，以及对取证结果的记录存档。

### 6．舆论导向

舆论导向针对敏感话题进行正面的舆论引导，尽量减少不良或负面舆论的影响。其主要功能包括自动回复和人工干预。

### 5.4.4 舆情呈现子系统

舆情呈现子系统架构图如图 5-9 所示。

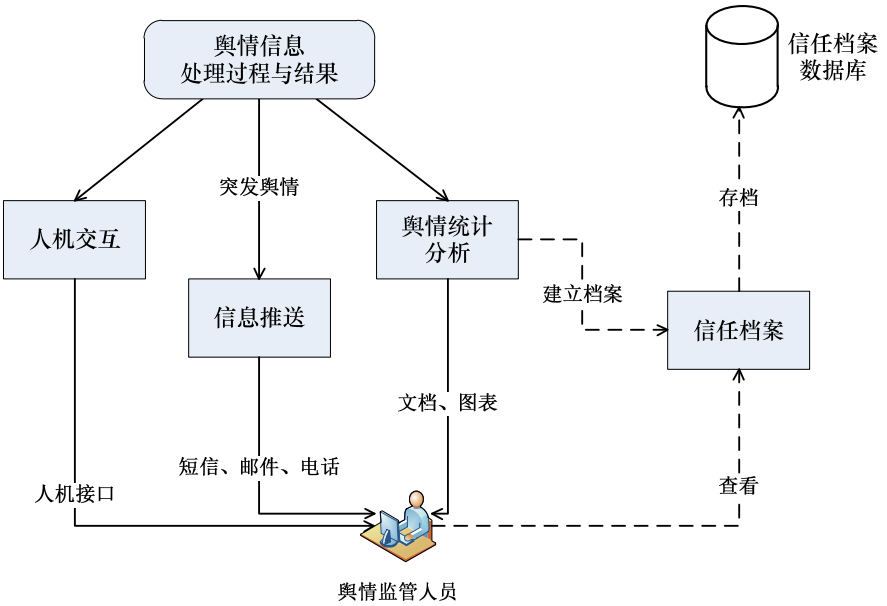



图 5-9 舆情呈现子系统——系统架构

#### 1. 人机交互

人机交互的作用主要体现在用户能够通过人机交互模块对系统的某些参数进行初始化，同时能够以友好的方式将用户需要的舆情服务返回给用户；用户能以可视化的方式对整个系统进行管理，也可以按照分类规律清晰地查看处理层处理得到的舆情信息。人机交互模块包括用户接口和动态显示。其中用户接口实现系统与用户之间的人机交互，动态显示则根据动态监控返回的监控信息进行实时反馈显示。

#### 2. 信息推送

此智能舆情监测系统的舆情呈现子系统提供信息推送的服务，该服务针对使

用手机、平板电脑等移动客户端的用户。由于经过舆情分析子系统和舆情处理子系统对数据进行相应的处理后，得到的舆情信息结果规模仍然相当庞大，本舆情监测系统采用信息推送的服务，有针对性地定期向特定的系统用户推送特定的舆情信息，做到有的放矢，帮助用户高效率地发掘有价值的信息，节约用户的时间。

向用户进行信息推送的依据主要有两个方面：一个方面是用户感兴趣的舆情类别，这是用户使用系统时人为设置的；另一方面是基于舆情处理子系统产生的预警信息。

### 3．舆情统计分析

舆情统计模块将得到的舆情处理信息进行分类显示，对相应分类下的各个舆情信息进行安全评级，也可根据不同舆情处理结果以合适的统计图表呈现结果数据，同时在原有处理结果的基础上分析挖掘出有价值的潜在规律并以统计图表的形式展现，最终以简报的形式推送给相关部门的舆情监管人员及高层决策者。该模块包含 3 个子模块，分别是分类评级、统计图表和常态舆情简报。

分类评级是指在舆情分析层中对 BBS 信息资源进行分析和处理后，将获得的网络舆情按不同类型进行分类呈现。同时在分析网络内容及舆情本身的性质和特点及舆情演化规律和条件的基础上，构建网络舆情安全评估指标体系，在分类显示不同网络舆情的同时显示相应网络舆情的安全等级。

统计图表以统计图或统计表的呈现方式，如表格、柱状图、曲线图、饼图等，来呈现舆情处理层所得到的全部处理结果。

常态舆情简报则通过舆情监控系统将采集的舆情信息自动分类生成舆情报告。系统根据从舆情分析子系统和舆情处理子系统中得到的信息，以日、周、月和年为单位，生成舆情简报。

### 4．信任档案

此系统信息采集子系统得到的信息，经过舆情分析子系统和舆情处理子系统

的处理得到的结果，可以帮助我们建立信任档案。信任档案包括两部分内容：用户信任档案和高校及网络社区信任档案。而信任档案的信息保存在数据库中，使用增量存储的方式，从而可以对高校 BBS 和网络社区进行长期监控及有效的备查。

### 5.4.5 统一管理平台

为了确保系统的所有子系统及模块可用、可信、安全、统一，构建一个统一的管理平台，用以实现对整个舆情管理系统全部资源的统一管理和调配，建立一个新的全局、智能的网络舆情管理体系。统一管理平台是舆情管理系统的管理与控制中心，它对系统用户进行管理权限配置，对各个子系统进行统一管理和配置，监测各个子系统的运行状态，收集管理系统日志，从而实现全网所有资源的集中管理。其主要包括 7 个部分，分别为：用户及权限管理、信任管理、系统资源管理、信息采集子系统管理、舆情分析子系统管理、舆情处理子系统管理及舆情呈现子系统管理，如图 5-10 所示。

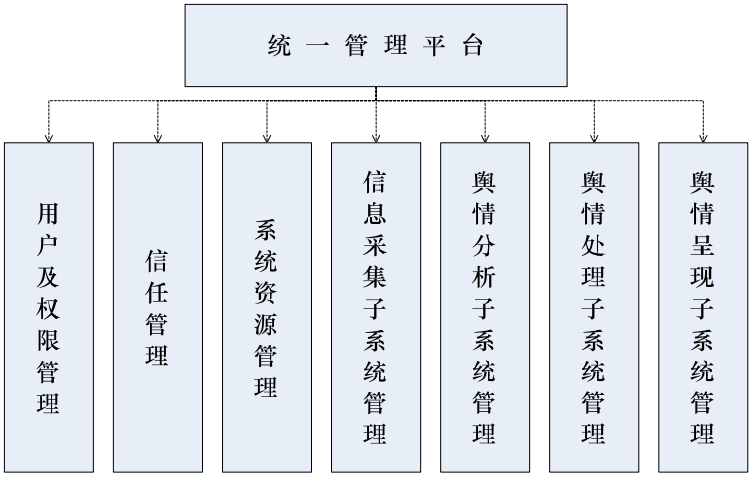



图 5-10　统一管理平台系统架构

### 1．用户及权限管理

身份认证与授权子系统，用来区分系统的登录用户及不同级别的用户组，对他们的身份和操作的合法性进行检查。

### 2．信任管理

通过分析用户在论坛参与过程中的各种网络行为，对论坛中的不同用户建立个性化的、独立的信任档案作为信任证据，为不同论坛各自的用户角色权限管理提供参考依据，实现对论坛用户的信任管理。

### 3．系统资源管理

本部分负责采集、维护系统服务器集群的所有信息，为风险管理、脆弱性管理提供分析依据。对本地的服务器集群作负载均衡，通过数据库资源管理，监控原始采集数据库、舆情数据库、舆情语料库、舆情策略库及信任档案库五大核心存储数据库的使用及容灾情况，保障数据库资源充足、负载均衡、合理分工，满足系统资源需求。

### 4．信息采集子系统管理

主要完成对信息采集子系统的相关模块控制参数配置，以实现监管人员的特定信息采集需求。

### 5．舆情分析子系统管理

主要完成对舆情分析子系统的相关模块控制参数配置，以实现监管人员的特定舆情分析需求。

### 6．舆情处理子系统管理

主要完成对舆情处理子系统的相关模块控制参数配置，以实现监管人员的特定舆情处理要求。

### 7．舆情呈现子系统管理

主要完成对舆情呈现子系统的相关模块控制参数配置，以实现监管人员的特定舆情呈现要求。

## 5.4.6　安全保障子系统

安全保障子系统的主要作用是保证用户访问及使用网络智能舆情监测系统时的安全，防止非法用户及黑客入侵监测，包括 eID 身份认证、SSL VPN 安全网关与网络安全，以及系统本身运行及数据的安全保障，包括日志审计与数据容灾。其架构图如图 5-11 所示。

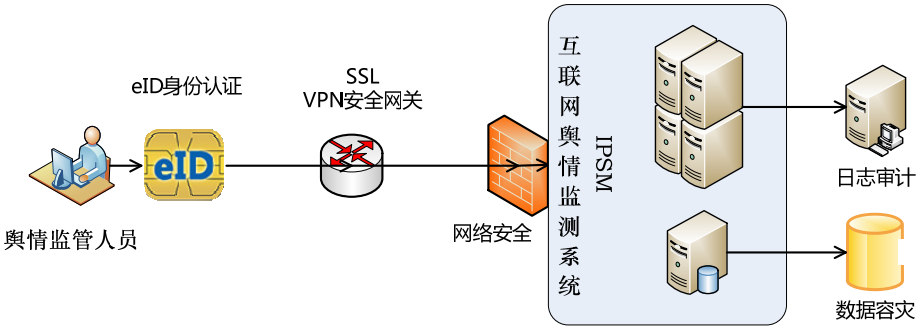



图 5-11　安全保障子系统——系统架构

### 1．eID 身份认证

网络电子身份证（Electronic Identity，eID）在网络空间可以唯一标识一个用户身份。eID 在网络上远程使用时，由 eID 载体的安全智能芯片进行密码运算，并由 eID 服务平台完成真实身份的验证识别。eID 可无缝嵌入网站及智能移动终端中，广泛适用于需要管理网络身份的电子政务、电子商务、金融支付、虚拟财富交易及社交网站等领域，以确保信息的私密性和安全性。使用 eID 身份认证可以有效地保证舆情监管人员的可信性。

### 2．SSL VPN 安全网关

SSL VPN 可以通过特殊的加密通信协议，在 Internet 上的用户与舆情监控中心之间建立一条专有的通信通道，就好比架设了一条专线。与传统 VPN 解决方案相比较，SSL VPN 使用和维护简单，不用更改现有的网络结构；移动性强，无须安全客户端程序；具有强有力的访问控制能力，可以使用户轻松访问舆情监控中心内部 B/S 和 C/S 应用及其他核心资源。

### 3．网络安全

整个系统的网络安全包括两个方面：网络边界的安全和网络内主机的安全。对于网络边界的安全，通过部署防火墙及入侵检测系统来确保网络边界的安全可靠性；对于网络内部主机层面的安全，采用立体化的防病毒体系，在工作站、服务器上安装相应的防病毒软件，由中央控制台统一控制和管理，实现全网统一防病毒。

### 4．日志审计

本模块提供系统平台运行日志、系统审计日志、网络操作日志的统一管理，系统审计员具有操作权限，包括日志的查询、统计、删除、备份和报表管理。

### 5．数据容灾

系统数据安全的目标是最大限度地确保数据安全和业务的连续性。本系统采用以下方法保障关键业务数据安全：备份关键数据、建立权限、对敏感数据加密。

## 5.4.7　主要技术指标

本系统采用的主要技术指标有如下几个。

（1）舆情获知范围：全国至少 50 个论坛，包含重点高校 BBS 及热门网络社

区，支持 IPv6、IPv4 协议。

（2）舆情获知平均时间：小于 10 分钟。

（3）舆情控制范围：突发舆情预警、URL 实时过滤、信息跟踪与取证、舆论导向。

（4）数据库指标：舆情语料库、舆情策略库自学习、自更新；信任档案库包含 10 万个记录用户档案信息。

（5）安全指标：支持 eID 证书、PKI 技术和 SSL 访问等。

## 5.5 其他舆情监测系统介绍

### 5.5.1 人民网舆情系统

人民网舆情监测室[2]是国内最早从事互联网舆情监测、研究的专业机构之一，在舆情监测和分析研究领域处于国内领先地位。人民网舆情监测室研发并完善了具备个性化、垂直性监测功能的互联网舆情监测系统。该系统基于网络舆情传播规律，及时、全面地监测境内外新闻网站、论坛、报刊、电视、广播和知名博客、微博，并在此基础上进行数据的抓取、挖掘、聚类、分析和研判，方便舆情工作人员迅速获取舆情，提高舆情管理和舆论引导的水平。

### 5.5.2 拓尔思

北京拓尔思信息技术有限公司[3]建立了 TRS 大数据舆情分析平台（TRS SMAS），此平台是基于云服务模式的互联网舆情分析服务平台。主要产品线：RS 互联网舆情管理系统（TRS OM），此系统通过互联网信息采集和文本挖掘技术，

帮助各级政府快速发现和收集所需的社会网络舆情信息，通过自动采集、自动分类、智能过滤、自动聚类、主题检测和统计分析，实现社会热点话题、突发事件、重大案情的快速识别和定向追踪，从而帮助政府及时掌握舆情动向，对有较大影响的重要事件快速发现、快速处理，从正面引导舆论和宣传，构建积极向上的主流舆论，并为政府决策提供信息依据。

### 5.5.3　鹰击系统

鹰击系统是国防科技大学计算机学院与湖南蚁坊软件公司合作研发的舆情监测系统[4]，其主要功能如下。（1）实时全面监测：每日采集 2 亿条微博，海量数据实时推送。及时将预警信息以在线提示、短信、邮件方式告警，离线也能及时应对突发事件。（2）深入直观分析：针对重点人物、群体、社交关系、事件发展趋势、言论倾向、传播路径等进行深入分析，为舆情“早响应”提供有力辅助。（3）快速响应互动跨、平台多账号管理，及时互动，快速回应网民关切，提升公信力。

### 5.5.4　Buzzlogic

Buzzlogic[5]是一家基于数据分析技术，从事网络广告制作、网络舆情分析、市场营销推广及企业公关策划的公司，其提供的“BuzzLogic Insights”服务通过对博客进行高时效的、全方位的、多角度的舆情动态分析，为营销人员提供产品反馈意见、品牌认知度情况；为公关人员提供与知名博客建立关系、发现并跟踪新舆情的服务；帮助企业发现、吸引及评估行业影响力，了解消费者需求，以改进服务。

### 5.5.5　Nielsen

尼尔森（Nielsen）公司是全球性的信息和媒体公司，它拥有领先的市场地位、

全面的媒介资讯，是出版界、展览界和报纸界公认的品牌。其提供的“BuzzMetrics”服务在全球口碑测量领域享有很高的声誉，它在结合经验、数据、技术的基础上，帮助企业对在线言论及传播行为进行分析，其中包括微软、福特、诺基亚、宝洁、索尼等全球知名企业，以增强企业在产品、市场、营销方面的竞争力，提升企业品牌形象，促进业务增长。

### 5.5.6 Reputation Defender

众所周知，互联网从根本上改变了隐私的概念。博客、微博、论坛及社会媒体的扩散创造了一个全球化信息交流的空间。互联网的增长、网络的特性、现实的状况使得管理网络声誉尤为重要。Reputation Defender[6]通过专有技术，帮助客户监控网络，删除负面舆论（服务的层次取决于收费的高低），为企业塑造良好的网络形象。如今，它已经为全球超过 100 个国家和地区提供过服务。

### 5.5.7 Visible Technologies

Visible Technologies[7]是一家从事网络品牌管理、网络营销推广及通信业务的公司，可帮助企业跟踪消费者舆情，管理相关搜索引擎，尤其是其提供的“TruCast”和“TruView”服务能为企业提供及时、全面、高效的战略解决方案，保护和促进企业的网络声誉。谷歌、雅虎、博雅、恒美、WPP 集团等都与之有过合作。

### 5.5.8 Cision

Cision[8]通过对博客、论坛等媒体网站进行大范围网络舆情监测，为客户提供全面的媒体资讯智能服务。公司的一站式综合解决方案致力于简化公司、公关代理、政府和非营利机构的公关活动，并将跨越两个主要的媒体领域：媒体测量和

资源分析，以帮助企业扩大覆盖范围、了解行业趋势、树立品牌形象及提高整体的公关和媒体监察能力。

## 5.6　本章小结

随着社交网络在社会舆情探索方面所显露出的突出作用，商业公司和国家对相关领域的研究也开始逐步重视起来。

本章简单介绍了网络舆情监测系统的实现，并介绍了国内外网络舆情的相关技术手段。在舆情研究方面也有许多工作可做，例如寻找更有效率、准确度更高的信息采集手段，加深对情感分析部分的研究等。

## 参考文献

[1]　蒋官宏．社交网络事件监测系统的设计与实现[D]．北京邮电大学，2015.

[2]　人民网舆情监测室概况，舆情频道（http://yq.people.com.cn/service/index.html）.

[3]　北京拓尔思信息技术股份有限公司（http://www.trs.com.cn/）.

[4]　蚁坊软件（http://www.eefung.com/）.

[5]　https://www.douban.com/note/130367686/.

[6]　https://www.reputationdefender.com/reputation/how-reputation-management-work

[7]　https://www.trucast.net/Login.aspx.

[8]　http://www.cision.com/us/.

# 第 6 章　品牌推荐与保护

## 6.1　引言

随着互联网的高速发展，消费者的消费习惯也发生了很大的变化，从传统的面对面消费模式，更多地切换到了省时省力的互联网消费模式。但不变的是消费者会为了保障自身的利益去寻求其他人的消费意见，以避免不必要的浪费；并且由于互联网使用的方便性和资源的普遍性，商品口碑的传播也会变得更为快速和广泛。在这种情况下，一个品牌的网络口碑会对其最终的销售利益产生致命的影响。品牌间的不良竞争在所难免，无良商家可能会利用消费者的这一消费习惯大肆对对手进行污蔑诋毁，降低其网络口碑和认可度，那么这时商家的自我保护尤为重要，很可能直接关系到品牌的生死存亡。本章探讨如何利用计算机技术来对品牌的网络口碑进行保护[6]。

## 6.2　网络口碑营销与网络水军

研究人员发现，人与人之间的沟通交流和信息共享严重影响着他们的偏好和抉择，也就是说一个物品、一个人或者一件事情的口碑决定着它们是否会被人们所接受和喜爱[1]。随着互联网的发展，以及 Web 2.0 时代的到来，产生了一种新的口碑模式，也就是网络口碑。由于人们的生活习惯和消费习惯很大程度上从传统模式转变成了互联网模式，近年来层出不穷的各类电商网站、商品评论网站、微博等都开辟了供用户畅所欲言的评论区，用户可以任意地对事物发表评论、表达自己的观点。某公司的调查报告显示，有 3/4 的消费者在网络购物前会去查看商品评论区的经验信息，甚至有 1/3 的用户会在实体店购买商品前在线查找有关商品的用户评论作为参考[2]。线上商品评论可以被任意用户浏览查看，这种信息的共享模式不受时间、地点、人数的限制，并且不会造成传播过程中的信息流失。因此，网络口碑可以影响的消费者人数远远超过传统口碑[3]。图 6-1 反映了传统口碑到网络口碑传播模式的变迁。

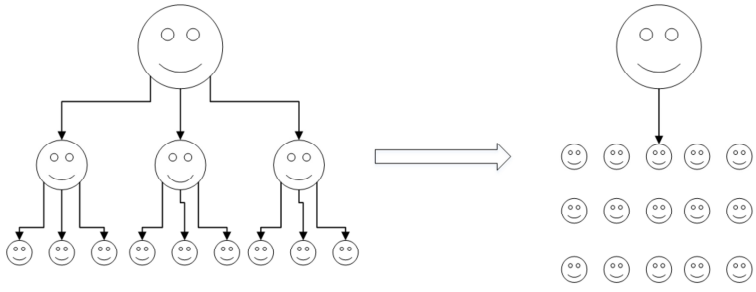


图 6-1　网络口碑模式变化

网络口碑的发展给人们的生活带来了极大的方便，很大程度地保障了消费者的利益。根据商品的网络评论，消费者可以全方位地获知有关商品的所有信息，清晰地知道该商品是否满足自身要求，从而快速做出是否购买的决策。然而近年

来网络口碑的参考价值却有大幅的降低，有些商家发现自家商品明明在各方面都具备优势，却因网络口碑的拖累导致销量低迷。而造成网络口碑不符实情的原因主要是一些黑心商家利用网络口碑的巨大影响力，采用一些不法手段来打击、污蔑竞争对手，具体做法就是花少量的钱雇佣大量的“托”，即“网络水军”[5]，假装成已消费的普通用户在对手品牌商品评论区域发表言论。这些言论通常带有污蔑、诋毁、谩骂的性质，由于这些言论数量的庞大，品牌的网络口碑自然会受到很大的影响，而这种影响导致的直接后果就是该品牌商品销量的大幅下跌。这种手段不仅仅用于电商网站商家之间的销量之争，也用来影响电视节目的收视率、电影的票房、明星等知名人物的公众形象等。下面几个例子显示了借助网络口碑的不正当竞争手段对上述几种类型的品牌所采取的具体行为及给它们带来的影响。

（1）在电影业，很多电影都经历过“网络水军”抹黑事件，最知名的莫过于陆川导演的《王的盛宴》。影片上映后，在一些电影评论网站上的评分一路下跌[4]，在豆瓣上甚至一度跌破 5.8 分，网友留言区尽是侮辱、谩骂的字眼和负面的声音，差评数达到 9000 多条。电影亏本几千万元之多，除了电影内容和各方面自身因素外，最大的外因就是遭网络“水军”抹黑，影响了排片量和票房。同样的情况在电影业屡见不鲜，严重扰乱了业界秩序。

（2）越来越多的用户在电子商务平台，如淘宝、京东、美团、大众点评等，选择中意的商家进行消费，选择的标准通常离不开这个商家的人气、销量及买家留言，这很容易被不正当对手利用，雇佣“网络水军”对商品进行诋毁。甚至还有“淘宝职业差评师”，他们受雇于商家甲，扮成普通网民到商家乙进行购物交易，然后给其差评，导致商家乙的商品销量大大降低。更有甚者，一些“职业差评师”团队并非受雇于他人，而是自发地以此手段威胁商家，讹诈钱财。

（3）明星因其职业的特殊性，常常会成为人们关注的焦点。有些当红明星会因为风头过盛而引来同行的嫉妒，遭受不正当打压。通常采取的手段是雇佣专业

团队在微博等明星社交网站评论区发表不实言论，伪装成明星粉丝，大肆发表褒扬该明星、贬低甚至辱骂其他明星的言论，造成普通民众的反感；或者直接发表诋毁、侮辱、对明星名誉道德有损的假消息，引导舆论方向，让民众觉得该明星人品有问题。

由于网络水军的匿名化，使其很难在短期内得到根治，口碑在现阶段主要靠自己来保护。评论数量和评论的情感倾向是影响用户消费决策的重要参数[8,9]。在商品受到大量不法评论攻击时，正面评论是品牌保护中十分重要的因素[10]，以回帖的方式对他们的评论进行澄清，以大量正面符合事实的言论对他们的诋毁言论进行淹没，在增加评论数量的同时使评论情感倾向趋于正面。

然而，以传统方式人为进行回帖反击是行不通的，因为品牌商家内部可以参与该活动的人数与恶意评论的“网络水军”数量相比存在巨大差距，正面评论的效率必然大大落后于负面评论的效率。而且“网络水军”往往专职于该工作，而品牌商家并没有很多时间可以和“网络水军”相互对抗，难以淹没“水军”的言论。

我们考虑搭建品牌推荐与保护系统，让程序自动完成烦琐的回帖过程，实现高效的批量回帖，而商家只需在操作界面完成简单的回帖配置操作，从而节省时间和精力。通过该系统，增加正面评论数量变得十分简单，网络口碑可以得到很好的保护。

## 6.3　品牌推荐与保护关键技术

品牌推荐与保护主要涉及三项关键技术，即评论采集技术、自动评论技术和验证码识别技术。下面主要介绍前两项技术，验证码识别技术将在第 7 章介绍。

## 6.3.1 评论采集技术

### 1. 信息爬取

想要自动保护品牌的网络口碑，首先要通过网络爬虫获知品牌被诋毁的具体言论、数量、时间、被何人诋毁等信息。网络爬虫的原理是从一个初始网页开始不断把新的 URL 放入待抓取队列中，直到队列为空 [16]。一个爬虫程序可将互联网页面分为如图 6-2 所示的 5 种类型[17]。

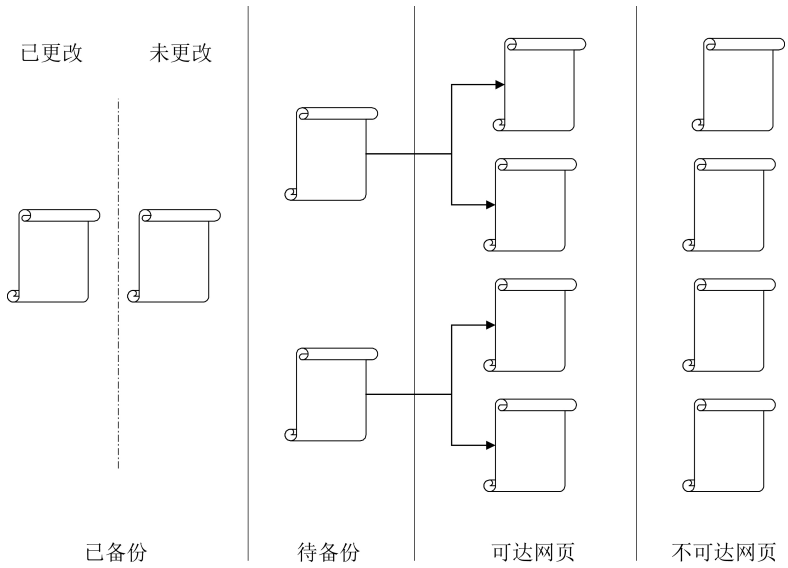



图 6-2 爬虫网页类型

这 5 种类型分别是已备份-已更改、已备份-未更改、待备份、可达网页和不可达网页。

（1）已备份-已更改页面，即该品牌商品页面已被爬取备份，但目前内容有所更改，如有新的评论出现。及时发现新的恶意评论对品牌保护来说至关重要，故需要为爬虫程序设置定时间隔，实时更新获知内容是否已发生改变。

（2）已备份-未更改页面，即该品牌商品页面已被爬取备份，并且目前内容未有更改。

（3）待备份页面，即存在于待抓取队列中的页面。

（4）可达页面，即尚未放入待抓取队列或已抓取队列，但通过待抓取队列中的 URL 可被放入待抓取队列的页面。

（5）不可达页面，即直到爬虫程序退出也不会被放入任何队列的页面。

上述页面分类反映了爬虫程序运行过程中页面状态的变化。其中在品牌保护技术研究中值得注意的是，已备份页面从未更改状态变为已更改状态表示很可能出现了新的攻击，需要实时关注；而不可达页面属于与品牌无关的页面，是不需要关注的页面。

### 2．主题网络爬虫

为了减少信息爬取量，我们采用主题网络爬虫[18]。主题网络爬虫只关心与主题相关的页面，所以计算页面的相似度以判断该页面是否需要爬取成为主题网络爬虫区别于一般网络爬虫最重要的一点。其基本原理是通过一定的相似度计算算法得出当前网页与主题的相关程度，如果计算结果大于预先设定的阈值，则该网页需要爬取；否则可以舍弃。在品牌自动保护中，我们设定的主题是拟保护品牌，希望爬取的页面是所有该品牌旗下的商品页面，包括各商品的评论详情。

主题网络爬虫是普通网络爬虫的特例，其主要特点在于有效地剪枝以避免资源和效率的浪费。所以主题网络爬虫中最重要的是采取合适的搜索策略，使用有效的主题相似度计算方法对要爬取的网页进行过滤，从而达到剪枝的效果。常见的策略和算法包括 Best First Search 算法[19]和鱼群算法。

### 3．网页索引

采集到信息后，还需要为各商品评论网页建立索引，提供信息的搜索入口。

为网页建立索引的原理是从爬取出的网页中选取索引项，用这些索引项代表该网页，并将它们存入预先建立的索引表中。通过这些工作，用户可以更高效地检索网页。

为网页建立索引[20]需要一个词典抽取网页中的索引项，词典包含各种形式的词语及词语 ID，词典主要是为了根据词语得到词语 ID。索引器用于计算各索引项在网页中的权值（通过出现次数计算），再将该权值及词典中的词语 ID 存入索引表。索引表是一个由词语 ID 和词语出现记录组成的文档。图 6-3 表示了一个完整的网页索引过程。我们使用开源的全文检索工具包——Lucene[21]来为本系统构建索引。

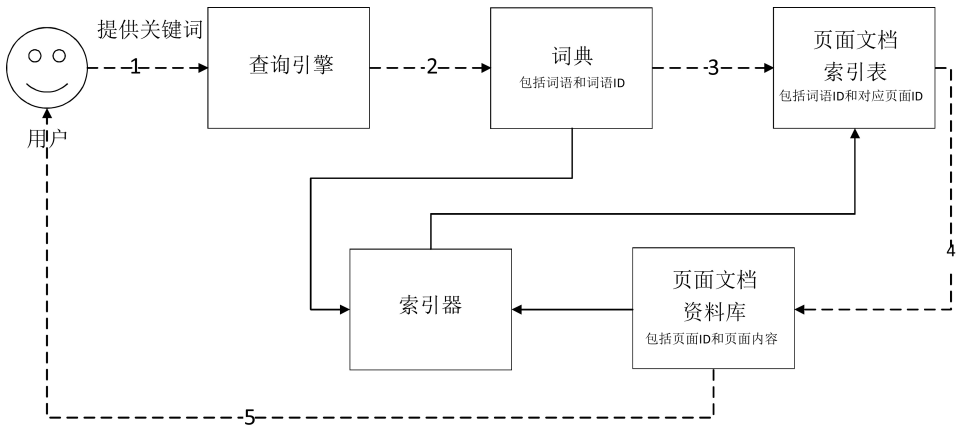



图 6-3　网页索引过程

### 4．信息结构化抽取

网页爬取下来后，信息通常并不是结构化的，还存在大量冗余信息，故必须用信息结构化抽取技术从这些非结构化信息中抽取出所需的结构化信息。例如，可以从电影评论网站爬取的网页信息中抽取出所需的网站名称、影评 URL、影评标题、电影评分、评论数量、评论用户名、评论时间、评论具体内容等信息。

现有的网页信息结构化抽取技术有很多，常用的包括基于视觉的信息结构化抽取技术及基于分装器的信息结构化抽取技术[22]。

（1）基于视觉的信息结构化抽取技术：所要抽取的信息在视觉上有一定的特征，利用信息间的分隔符来识别有用信息和无用信息，从而对目标信息进行抽取。

（2）基于分装器的信息结构化抽取技术：首先建立一个模板库；然后从一些网页中抽取框架作为该网页的基础模板；接着利用机器学习技术，使用很多其他网页对该模板进行改进；最后从模板库中选取相应模板与目标网页进行对比，抽取出结构化信息。

### 6.3.2　自动评论技术

自动评论技术是品牌推荐与保护系统的核心，利用自动评论技术批量、高效地对诋毁进行还击和淹没，达到商品品牌保护的效果。本节介绍基于 HTTP 协议的自动回帖技术和评论情感倾向性分析等。

评论通常以回帖的形式进行，回帖可以在品牌网站的具体某一个商品评论页面的评论区域中进行。通常情况下，回帖这一功能都是通过浏览商品具体页面—进入相应评论页面—编辑内容—点击发送等过程来完成的，而有一些必须要登录才可以回帖的网站则在回帖前还要有一个登录网站的过程。自动评论技术实将这些过程封装在程序中，而不需要用户过多参与。

我们用程序模拟在浏览器中回帖的过程，实现自动回帖。自动回帖的原理是：首先理解回帖中涉及的 HTTP 协议是如何运作的，然后提炼出影响回帖各阶段的 HTTP 协议参数，最后转化为目标语言的程序代码。客户端向服务器端发送的 HTTP 请求[26]中需要指明请求方法，可以是 Get、Post、Head、Put、Delete、Trace、Connect、Options 等，其中最常用的是 Get 和 Post 方法。

Get 方法一般用于从服务器端获取资源信息，不会修改资源，但是使用 Post 方法必须在表单中完成。因此，为了寻求方便，很多人会用 Get 方法来提交信息给服务器，方式通常是将需要提交的信息以键值对的形式附在 URL 后面，以 URL 参数的形式传递到服务器端，用“？”将信息与 URL 隔开，多个信息之间用“&”间隔。由于浏览器对 URL 是有长度限制的，所以可想而知 URL 参数的个数也是有一定限制的。用 Get 方法发送数据的安全性并不高，因为这些数据以 URL 参数的形式附加在 URL 后面，所以会完全暴露在用户面前。在传递用户名、密码等重要信息时，从安全的角度考虑，不建议使用 Get 方法。

Post 方法通常用于发送数据给服务器端以更新服务器端资源，它会修改资源。Post 方法传递的数据被存放于 HTTP 包的包体中，没有长度限制，因而可以发送大量数据。并且因为信息都封装在包体中，不会暴露给用户，因此安全性很高。

可以通过采用浏览器提供的开发者工具进行抓包，通过分析一次发帖涉及的多次交互及数据交换找出关键的中间参数。还需要注意网站是否需要登录的问题。系统面向的电商、电影、娱乐等网站分为可以匿名回帖的网站和不可以匿名回帖的网站，对于不可匿名回帖的网站则必须要有一个登录的过程，即需要多一次 Post 操作来向服务器提交用户名和密码。这里用 Post 操作而不用 Get 操作的原因是 Get 操作将用户名、密码以 Query String Parameters 的形式附加在 URL 串后面，容易暴露信息，是不安全的。

综上所述，品牌网站按照可否匿名回帖可以分为两种类型，分别是可以匿名回帖的网站和不可以匿名回帖的网站。可以匿名回帖的网站的后台回帖过程只需找到相应回帖页面，提交回帖内容即可，即只需一次 Post 请求即可；不可以匿名回帖的网站的后台回帖过程由登录+回帖两个步骤组成，需要提交用户名、密码、回帖内容等参数，需要两次 Post 请求。然而登录+回帖这个过程并不是两次 Post 过程的简单相加，因为在登录过程中，第一次 Post 过程完成后的状态对第二次 Post 过程有着重要的影响，也就是说只有在登录成功后才能进行回帖，所以登录

+回帖的全过程实际上应该是 Post+状态记录+Post 的过程。HTTP 协议采用 Cookie 来提供状态记录的场所。在完成登录操作后，我们可以从收到的响应头中提取 Cookie，再将其放入回帖操作请求头中的 Cookie 中。图 6-4 以某不可匿名评论的网站实例表示该过程。


图 6-4　回帖实例

其中的所有参数都可以通过抓包获取。登录 Post 的 Request Header 部分待定是因为通过抓包观察到的请求头部非常详细，我们完成自动回帖操作可以把所有头部信息都写到程序当中。但并不是所有数据都是必需的，出于节省效率和资源的考虑，经模拟排查分析，基本上所有的 HTTP 请求 Header 都不超出 Content-Type、Referer、Cookie 这三个字段。登录 Post 因为没有前置的 HTTP 请求操作，所以不需要指定 Cookie。在提交登录 Post 请求后，我们会收到服务器的回复，即 Response Header 和 Response Body。Response Header 中记录此次请求完成后的状态信息，其中我们需要的是名为 Set-Cookie 的若干条字段，这些就是回帖 Post 需要的全局登录状态信息。

经过以上分析，基于 HTTP 协议的自动回帖基本原理已经很清楚，只要将这些原理转化为程序代码，就可以完成品牌推荐与保护最核心的部分。具体实现可以采用第三方.jar 包 HttpClient 进行 Java 网络编程。HttpClient 是 Apache Jakarta Common 的子项目，它是一个支持 HTTP 协议的编程工具包。

在本系统中，由于电商网站、电影评论网站、娱乐网站等不同种类的网站中回帖内部参数不尽相同，故无法写一套程序来完成所有网站的自动回帖功能。为

了提高回帖成功率，我们针对每个网站单独编写回帖程序，再集合到一个总类中。如图 6-5 所示为某网站中的自动回帖流程图。

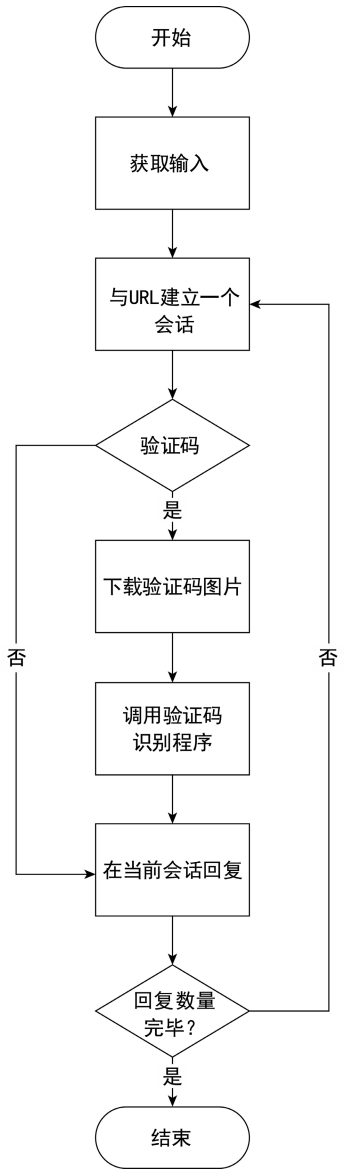



图 6-5　自动回帖流程图

### 6.3.3　评论情感倾向性分析

人们在发表言论的时候往往带有某种情感倾向，即言论中的观点和态度对于某个事物持积极态度、消极态度或中立态度。商品评论区是供人们任意发表对商品看法的地方，并没有任何限制，也就是说任何人都可以在这里发表意见，那么这里就必定存在各种情感倾向，有积极正面的，有中立的，也有诋毁和侮辱的。我们需要反击淹没的只是那些恶意的言论，积极和中立的观点在本系统中并不在我们关心的范围内，所以需要采用评论的情感倾向性分析技术，从采集到的评论信息中挑选出需要反击的评论打上标签，有助于后期筛选，提高品牌保护效率。

情感倾向性分析的方法有很多种，基于我们所针对的文本类型，即商品评论的特点——短小、简练、没有太复杂的情感词汇，本文采用简单的、基于词典迭代的情感倾向分析方法[34]。该方法的主要原理是不断迭代扩充情感词典，根据词典中的情感词逐层判断评论中短句的情感倾向性、长句的情感倾向性及评论全文的情感倾向性。完成这个过程需要三个步骤，分别为评论句子划分、评论情感倾向分析及情感词典更新。

#### 1．评论句子划分

在初始状态下，情感词典中只有一些基础的情感词，我们需要从评论素材中选取情感词放入词典中，迭代更新词典。情感词是那些在评论中出现两次以上的词。要统计评论中词语的出现次数，必须先对评论全文进行分词处理。又由于最后的目的是判断评论全文的情感倾向，我们采取的方法是逐层判断短句、长句、全文的情感倾向，所以还需要将评论全文划分成长句和短句。

- 分长句。划分的标准是以一些表示句子结束的标点符号作为分隔符进行长句分割，如中英文环境下的句号、感叹号、问号、分号、冒号等。
- 分短句。一个完整的句子往往会很长，我们需要将其分成若干短句。短句

往往是一连串连续的、不包括标点符号的汉字，所以划分短句的分隔符除了长句划分中那些句子结束型标点符号外，还包括逗号等句中间歇行的标点符号。

- 分词。词语的分割除了所有标点符号外，还需要一些字符和汉字作为分隔符，这些字符和汉字的选择非常重要。非汉字的数字、字母、空格、符号等可以作为分隔符，因为通常这些字符会插在一些词语的间隔中，而不会出现在一个词语的中间；“的”、“是”等汉字可以作为分隔符，因为这些词通常词性为助词，在语句中并没有实际的意义。

如图 6-6 所示为某电商网站中一条针对某品牌羽绒服的评论。

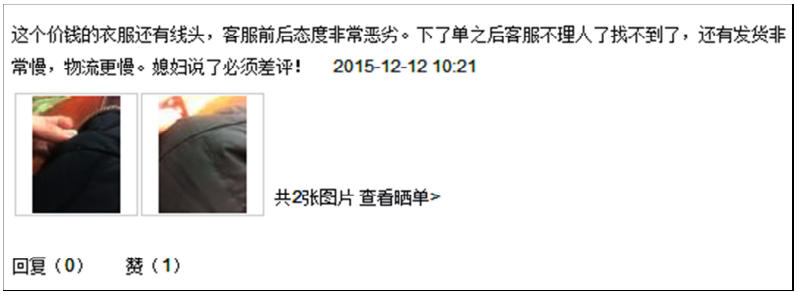



图 6-6　某电商网站商品评论

对该评论分长句、分短句、分词的结果如下。

- 分长句：这个价钱的衣服还有线头，客服前后态度非常恶劣；下了单之后客服不理人了找不到了，还有发货非常慢，物流更慢；媳妇说了必须差评。共有 3 个长句。
- 分短句：这个价钱的衣服还有线头；客服前后态度非常恶劣；下了单之后客服不理人了找不到了；还有发货非常慢；物流更慢；媳妇说了必须差评。共有 6 个短句。
- 分词：价钱；衣服；线头；客服；态度；恶劣；下；单；客服；不理人；找不到；发货；慢；物流；慢；媳妇说；差评。共有 17 个词语。

### 2．评论情感倾向性分析的方法

首先根据短句中词语的情感倾向性统计判断短句的情感倾向，然后根据长句中短句的情感倾向性统计判断长句的情感倾向，最后根据评论全文中长句的情感倾向性统计判断评论的情感倾向。

假设目前情感词典中只有“差评”、“恶劣”等负面情感词，以及“好评”、“棒”等正面情感词，对上文中已分词分句的电商网站评论实例进行情感分析的结果是：根据短句中正负面情感词的数量，在 6 个短句中有 4 个是负面短句，2 个是中立短句；3 个长句全部为负面长句，所以该条评论的情感倾向为负面属性。

### 3．情感词典更新

所有情感词的集合叫作情感词典[35]，这些情感词均存在一个情感指数，该指数介于–1～1 之间，反映情感词倾向于正面或负面的程度。指数越接近 0，情感越倾向于中立；指数越接近–1，情感越倾向于负面；指数越接近 1，情感越倾向于正面。情感词的情感指数由情感词出现的频率决定，计算公式如下：

$$D=\frac{|P-N|}{(P+N)/2} \tag{1}$$

其中，$D$ 表示该句中情感倾向的易区分程度，$P$ 表示正面情感词在句中出现的频率，$N$ 表示负面情感词在句中出现的频率。如果该值小于 1，则说明该句正负面倾向不是很明显，不建议作为情感词放入词典；如果该值大于等于 1，则说明该句正负面倾向明显，可以作为情感词放入情感词典中。其情感指数的计算公式如下：

$$\frac{P\text{或}N}{P+N} \tag{2}$$

其中，分子为 $P$ 还是 $N$ 取决于总体情感倾向趋向于正面还是负面，即如果 $P>N$，则说明情感倾向于正面，分子为 $P$；如果 $P<N$，则说明情感倾向于负面，分子为 $N$。

利用以上方式计算可以选出要放入情感词典的情感词及其情感指数，对情感词典进行更新。如果情感词典中已经存在该情感词，则更新其情感指数；否则直接将该情感词及它对应的情感指数放入情感词典中。

通过上述几个步骤可以很方便地完成对评论信息的情感倾向性分析工作，在信息存储的时候把每个评论的情感倾向性以标签的形式一起存入数据库，使用户可以方便地对具有某种情感倾向的信息进行查找，提高评论效率。

## 6.4 品牌推荐与保护系统

下面介绍我们自行开发的品牌推荐与保护系统。上一节介绍了其中涉及的关键技术，本节探讨如何将这些技术整合到一起形成可供用户操作的品牌保护系统。

### 6.4.1 系统架构

本系统采用 B/S 体系结构，即浏览器/服务器结构，这也是现今应用最为广泛的一种结构。在这种体系结构下，用户通过浏览器访问系统。相较于传统的 C/S 结构，B/S 结构的系统开发和维护得到了大大的简化，可移植性好，用户操作也变得更简单。如图 6-7 所示为 B/S 系统结构模型图。


图 6-7 B/S 系统结构模型图

### 1．总体架构设计

本系统分为两个独立的功能模块：信息采集处理与展示模块和自动评论模块。两个模块使用不同的数据库。信息采集处理与展示模块仅负责采集品牌商品相关信息，主要是评论信息，将它们进行处理后存入该模块的数据库中，并在系统中以一定格式展示，供用户查看，实时监测品牌被诋毁情况；自动评论模块负责在品牌网页进行实际的自动评论操作，包括多个子功能模块，供用户方便、高效地保护品牌的网络口碑。这两个功能模块彼此独立，互不影响工作，即使其中一个模块出现问题，也不会影响另一个模块的工作，但前者又在一定程度上对后者的工作有一定的帮助和效率上的提高。本系统总体采用数据、视图、功能逻辑分开的三层架构，图 6-8 为系统的总体架构图。

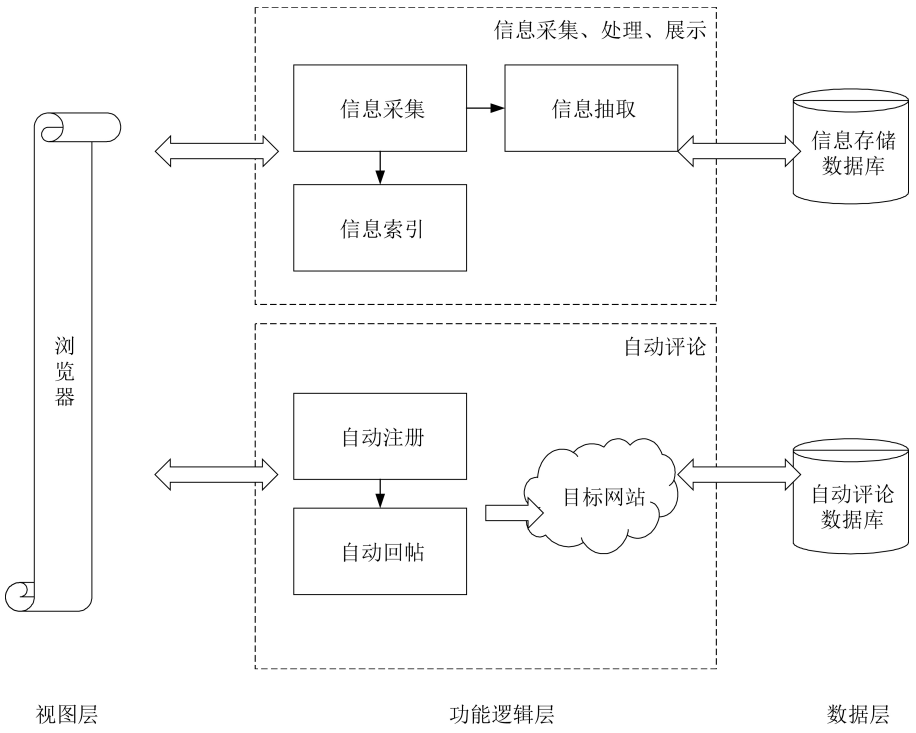



图 6-8　系统总体架构图

### 2．信息采集与展示模块架构设计

信息采集与展示模块采集抽取相关网站中品牌商品的评论信息，并建立索引，存入数据库，格式化展示在系统中。用户可以通过前端页面实时查看相关信息，也可以通过系统对信息进行筛选、搜索、关注等操作。信息采集使用 WebMagic 工具箱实现，它是一个简单灵活的爬虫框架；信息索引检索使用 Lucene 工具箱实现，它是一个全文检索工具包。信息采集与展示模块架构图如图 6-9 所示。

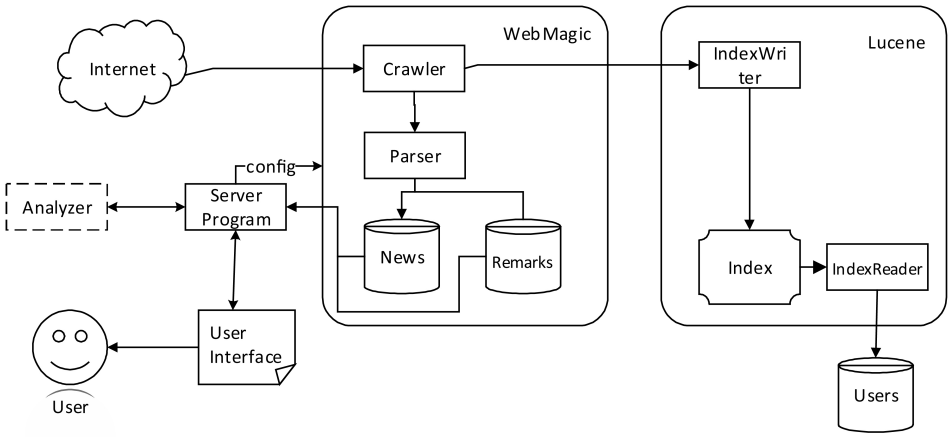



图 6-9　信息采集与展示模块架构图

信息采集与展示模块分为两个子模块：爬虫模块和 Web 模块。Web 模块基于 IndexReader 处理相应用户的请求；爬虫模块基于 WebMagic 和 IndexWriter，在启动时开启一个定时任务，定时间隔可以设置，启动爬虫，根据配置的入口页面列表爬取网页并添加到索引中。

### 3．自动评论模块架构设计

自动评论模块是品牌推荐与自动保护系统的核心，该模块通过对目标网站中的目标页面自动回帖完成品牌网络口碑保护的功能，用户只需通过自动评论模块

间接与目标网站交互，不需要进入目标网站。模块还提供自动回帖过程中需要的其他辅助功能。自动评论模块架构图如图 6-10 所示。

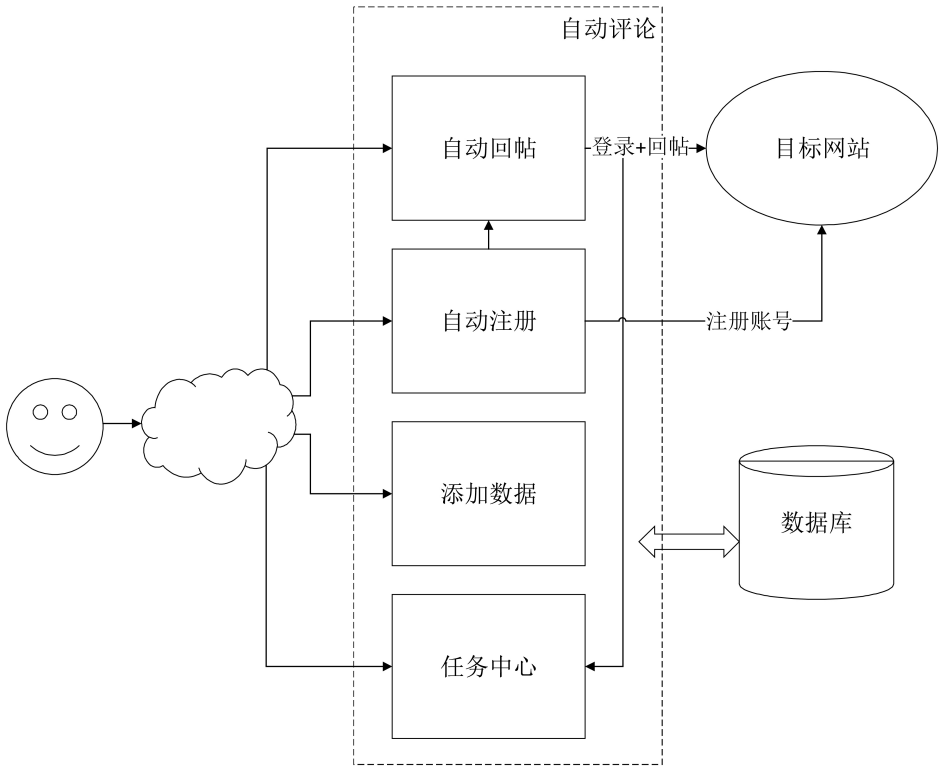



图 6-10　自动评论模块架构图

## 6.4.2　系统功能

通过系统架构分析可知，品牌推荐与自动保护系统分为两个独立的功能模块，分别为信息采集与展示模块和自动评论模块，自动评论模块又分为自动回帖、自动注册、添加数据、任务中心等几个子功能模块。系统功能模块图如图 6-11 所示。

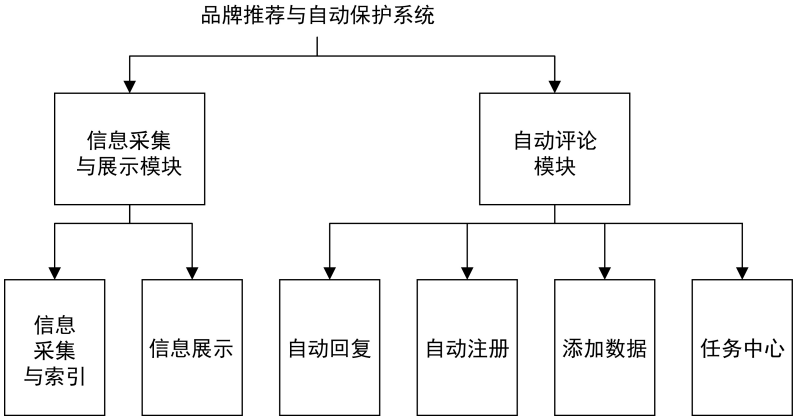



图 6-11　系统功能模块图

### 1．信息采集与展示模块功能设计

该模块供用户实时掌握品牌被诋毁情况。用户通过操作界面可以查看模块后台爬取到的品牌商品有关信息。以电影评论网站为例，用户可以在该系统中查看保护目标电影有关的所有影评及相关信息，包括网站名称、影评页面 URL、影评题目、影评作者、影评时间、影评原文、影评回帖数量、影评所有回帖内容及时间等；用户可以根据自己配置的关键字对这些内容进行搜索查看，找到自己关心的内容；用户可以配置自己的关注列表，若关注列表中的影评有新的评论出现，系统会予以提醒，用户可以第一时间获知是否有污蔑性言论出现。

由于目标网站具有一定的复杂性，含有大量的冗余信息，而我们的目标只是保护特定的品牌，故在信息采集模块中，在网络爬虫启动前，需要先对主题关键词进行配置，只爬取特定主题相关的信息，如电影评论网站的爬虫主题可以是电影名称。爬虫程序根据该主题关键词爬取信息，并建立索引，抽取有用信息，格式化存入数据库中。信息采集的流程如图 6-12 所示。

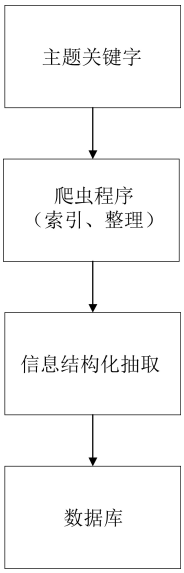



图 6-12　信息采集流程

在完成信息采集后，将相应数据表中的信息展示在系统中，并完成对信息的检索、关注、新消息提醒等功能，即可完成信息展示。

### 2．自动回帖模块功能设计

自动回帖模块是整个系统的核心，用户通过该模块完成自动评论的功能。用户可以将在信息采集与展示平台中查看到的回帖目标信息数据加入自动评论模块的数据库中，并且可以通过自动评论模块操作这些信息，完成自动评论功能。

用户使用自动回帖功能时，可以自己选择目标网页进行自动回帖，一次回帖可以选择一个或多个网页，并且网页必须属于同一个网站；用户可以自行选择回帖账号（可以匿名的网站中不必选择）、一次回帖的数量及回帖的素材进行回帖，账号及素材等数据均事先存储于数据库中；用户可以对目标网页进行搜索和条件筛选，并且可以自行在页面中添加目标网页，添加时只需输入目标网页的 URL，程序会自动提取其网页标题（以电影评论网站为例，则自动提取影评标题）；用户还可以删除已经废弃或太过久远的目标网页。

基于 HTTP 协议的自动回帖程序已经封装于后台，系统根据用户选择的目标网页选择调用相应的程序对请求进行处理。程序会自动判断在目标网页中回帖是否需要登录、判断登录的账号类型（用户名、密码、邮箱中需要哪些数据）、自动调用验证码识别程序、根据用户配置自动分配账号和回帖数量之间的关系，并给予回帖是否成功的反馈。如图 6-13 所示为自动回帖功能流程图。

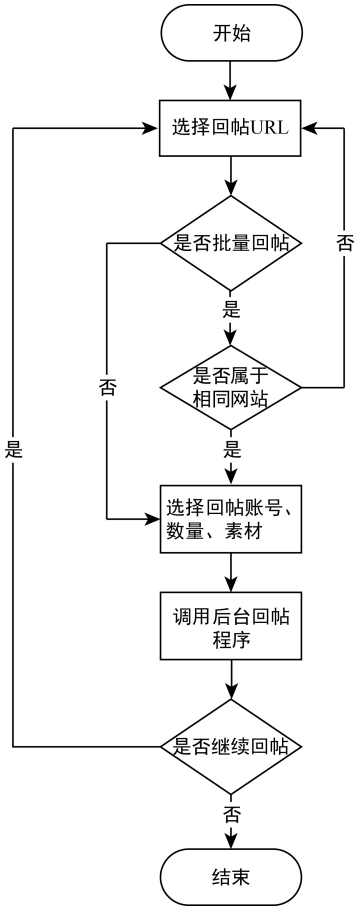



图 6-13　自动回帖功能流程

### 3．自动注册模块功能设计

由于目标网站通常不可以匿名回帖，对品牌进行诋毁的“网络水军”数量又

相当庞大，品牌商家仅仅依靠一人或几人的力量对品牌进行网络口碑的维护并不可行，商家给出的澄清回帖数量至少要达到与诋毁言论数量相当或几倍于诋毁言论数量，才能达到反击和淹没的效果。这可能需要几百甚至上千条正面回复内容，如果这些回复均出自一个或几个账号，则消费者并不一定会相信一家之言，则品牌保护不能实现，黑心商家则达成其诋毁打压对手的目的。为了防止上述情况发生，通过自动回帖来进行品牌保护需要大量目标网站的账号。通常注册账号步骤烦琐，需要填写的内容也多，不适宜用人工注册的方式来获取目标网站的账号，故本系统实现自动注册的功能。

用户可以在实现了自动注册功能的网站列表中选择目标网站，一键注册该网站账号。自动注册的后台功能实现原理与自动回帖类似，都是利用 HTTP 协议提交数据给目标网站来完成的，只是与自动回帖相比少了自动登录的过程。如图 6-14 所示为自动注册功能流程图。

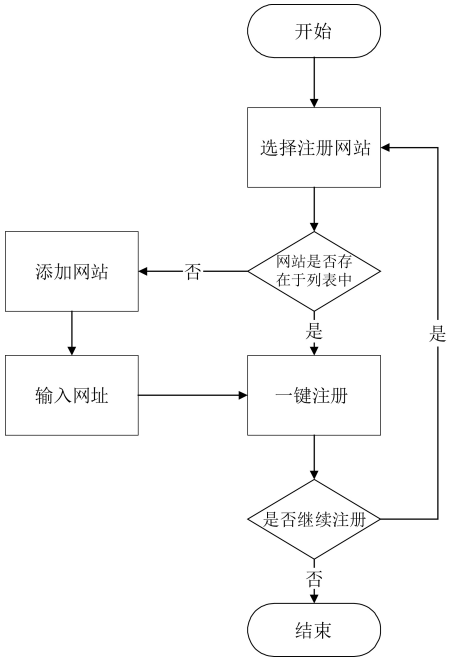



图 6-14　自动注册功能流程

### 4．添加数据模块功能设计

本系统中用于自动回帖的参数都由用户选择配置，但是品牌商家在使用该系统完成品牌保护时，可能会需要自行编辑信息，如一些有针对性的素材回复、已存在的目标网站账号等，并且这些新编辑的信息需要重复利用。最好的方式是将商家自己编辑的信息存入数据库中。但允许用户直接操作数据库并不是一个简单、安全的方法，所以本系统提供添加数据功能，用户只需在相应区域自行编辑内容进行保存，数据就会被存储起来供用户重复使用；用户也可以将想要导入系统的信息存成文件批量导入。如图 6-15 所示为添加数据功能流程图。

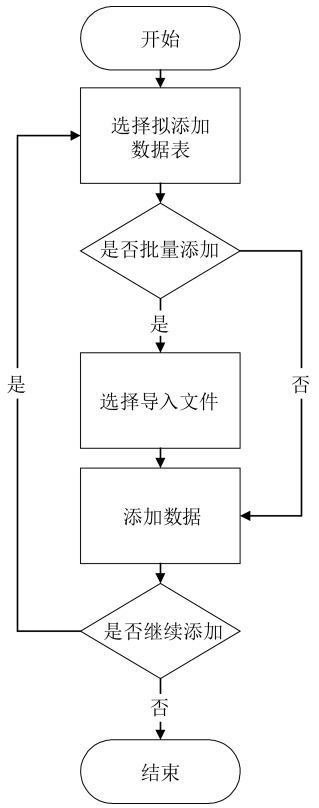



图 6-15　添加数据功能流程

5．任务中心模块功能设计

用户利用该系统进行对目标网页的批量回帖，每次回帖都需要等到后台程序把相应数量的帖子回复完，才能从系统页面上看到反馈，了解本次回帖是否成功。然而有时由于网络不稳定等因素，等待反馈的时间会很长，用户每次看到反馈确认结果，再去操作进行下一次回帖，会在等待的过程中浪费很多时间，造成回帖效率的下降。本系统实现了任务中心功能来解决这个问题。用户每次回帖后，回帖信息都会被马上存入任务中心当中，并以多线程的形式在后台运行。用户不必停留等待回帖反馈，可以继续进行下一次回帖，回帖结果可以去任务中心查看。如图 6-16 所示为任务中心功能流程图。

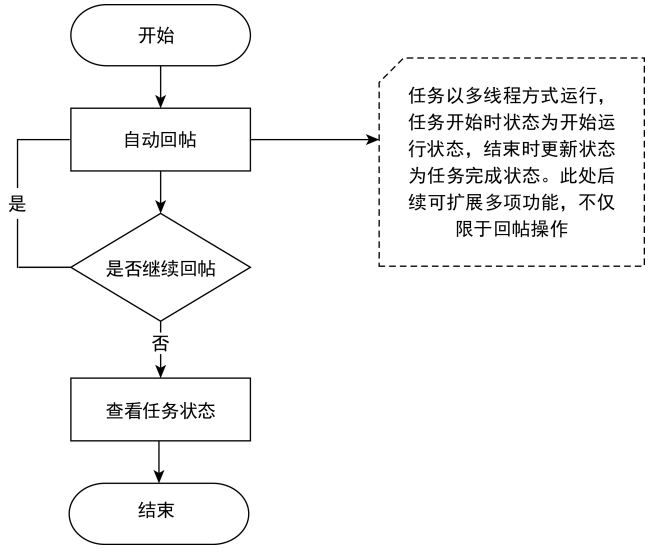



图 6-16　任务中心功能流程

## 6.4.3　系统数据存储

系统的数据存储采用两大功能模块分离的形式，即信息采集与展示模块和自动评论模块不共用一个数据库。这样的设计避免了两个模块间的相互影响，将其

中任何一个模块分离出来都可以作为一个独立的系统使用。

信息采集与展示模块的数据存储主要是将爬虫程序采集的信息经过结构化抽取后存入数据库中，供用户在展示平台中关注、查看；自动评论模块的数据存储主要是将在信息采集与展示模块中观察到的有用数据，以及一些自动回帖中需要的数据存入相应数据表中，供用户在自动回帖中使用。该系统设计的数据表包括信息采集与展示模块中的商品信息列表、评论列表，以及自动评论模块中的网站列表、情感列表、账号列表、素材列表、回帖页面列表、任务列表等。

## 6.5 网络水军识别研究现状

社会生活的高度信息化、巨大的用户群与潜在的商机，使虚假意见和垃圾信息被广泛地制造和传播，该类危害的源头即俗称的网络水军。网络水军形成巨大的虚假舆论场，影响网络民意、扰乱网络秩序、妨害经济利益，急需识别和治理。因此，网络水军识别关键技术已成为当前数据挖掘领域极为活跃的研究之一，本节主要介绍一些网络水军识别的技术。

### 6.5.1 网络水军识别简介

#### 1. 网络水军识别研究的基本概念及其特点

网络水军是指那些由商业利益驱动，为达到如影响网络民意、扰乱网络环境等不正当目的，通过操纵软件机器人或水军账号，在互联网中制造、传播虚假意见和垃圾信息等网络垃圾意见产生者的总称[7,11-15,23-24]。网络水军识别即在当前网络环境中，运用 Web 信息挖掘技术[25]，定义高区分度特征及行为模式发现潜藏的网络水军。网络水军也可以理解为整个网络用户中的离群点[27]，但其特征与正常用户十分相近，因此识别难度较高。

网络水军具备如下特点[11,28,29,30]：（1）目标相同，网络水军进行危害行为的目标大多是获得经济利益或造成网络影响；（2）数量巨大，网络水军为达到其目的，造成网络影响，必然会大量利用水军软件机器人（后文简称“水军机器人”）或傀儡账号；（3）行为异常，因其非正常动机，网络水军的行为模式区别于正常用户。

这些特点使得网络水军识别研究从统计角度具有可行性，为网络水军识别研究提供了基本的研究途径。网络环境的复杂和用户关注的增加，使水军行为模式隐蔽复杂化，并逐渐趋向于正常用户，也使得对其进行识别研究的难度加大。

### 2．传统网络水军识别研究

便捷邮件服务的流行，使互联网开始承载大量用户信息。在早期的网络环境中，获得用户邮箱和使用虚假邮箱的代价极小，初期邮件用户极易受影响，使得邮件领域网络水军泛滥。其运作方式主要表现为：通过大量发送垃圾邮件引导用户前往商业性质网站，或通过水军机器人[31]发布海量垃圾邮件，以最大限度地传播垃圾信息。

上文所述即传统网络水军，其出现时间较早、数量规模相对较小、行为没有高度隐蔽性，产生的垃圾信息具有明显特征。因此，对其的识别方法主要为基于垃圾信息内容分析，如邮件内容分析[32,33,36-39]。同时，通过大量识别建立黑名单和白名单分别用来记录可疑用户信息和正常用户信息，以此提高水军识别效率及准确率[32]。此外，邮件领域网络水军产生垃圾邮件所需资源类似，通过其使用资源及其网络层级特征能够很好地定位邮件水军。随着网络环境的复杂化和水军危害的增加，用户对其防范的能力也不断增强。为达到其目的，网络水军的行为逐渐复杂化并趋向于正常用户，传统邮件水军的识别方法无法发现这些隐蔽的网络水军。

### 3．Web 2.0 网络水军识别研究

Web 2.0 网络水军识别研究是对传统网络水军识别研究的延伸与扩展，是网

络环境变化衍生出的新型网络问题解决方案[40]。Web 2.0 网络水军识别研究难度加大，较传统网络水军识别研究面临更大的挑战。

## 6.5.2 网络水军识别的关键技术研究

Web 2.0 网络水军识别研究是传统网络水军识别基础上的适应性识别研究。目前，国内外网络水军识别研究取得了较前几年更大的进展，但是仍然存在很多重要问题亟待解决。国外网络水军识别研究最初集中于邮件领域，并在近几年内迅速扩展到社交网络和电子商务领域。国内网络水军识别研究相比之下较为缺乏。本文重点介绍基于内容特征、用户特征、环境特征及综合特征的网络水军识别方法，并对邮件、电子商务、社交网络、论坛等互联网重要应用领域内的网络水军识别研究关键技术进行了对比分析和总结。按照网络水军识别方法采用特征的不同，将网络水军识别方法分为基于内容特征、基于用户特征、基于环境特征和基于综合特征的识别[41]。

### 1．基于内容特征的网络水军识别研究

早期网络水军识别研究着重分析网络水军产生的内容，这是由于早期网络环境中网络水军产生的内容具有显著的可识别特征，如包含显著的商业广告信息和垃圾邮件信息；并且早期网络环境中用户防范性较差，该类网络水军能够造成的影响巨大。基于内容特征的网络水军识别研究涉及机器学习中的自然语言处理分支。该类含有观点的文本处理包括文本分类、文本情感分析及文本倾向性分析等方面。

### 2．基于用户特征的网络水军识别研究

随着网络环境逐渐复杂多样和用户辨别力的增强，使得制造传播具有显著特征内容的传统网络水军造成的影响不断降低。为了不断制造网络影响、妨害商业利益，网络水军逐渐衍生出多样的欺骗策略。其行为趋向于正常用户的行为，其发布内容也不再具有显著特征。通过分析变化的网络水军行为，基于用户特征的

识别研究能够很好地发现潜藏的网络水军。因此，当前网络水军识别研究转向基于用户特征的识别，以实现从源头遏制网络水军和垃圾信息泛滥的目的。此识别方法分为两种：基于用户行为特征的网络水军识别和基于用户关系特征的网络水军识别。

### 3．基于环境特征的网络水军识别研究

隐蔽的网络水军对用户展现出趋向于正常用户的特征，但其异常行为使其在网络环境层级表现出不同于正常用户的特点。Ramachandran 等人[40]在 2006 年首次提出基于水军网络级别特征的识别方法。他们从被水军污染的领域追踪收集了 17 个月共 1000 万条垃圾邮件信息，将该数据与基于 IP 的黑名单信息、TCP 脚印信息、路由信息及机器人网站命令追踪信息等联系起来，对水军的网络级别特征进行分析，实现垃圾邮件的追踪。此外，他们还提出了邮件水军在网络级别的危害策略。

基于环境特征的网络水军识别研究依据网络水军进行危害行为时产生的环境特征，该环境特征是无法被网络水军修改掩饰的，因此其识别准确率较高。但基于环境特征的网络水军识别研究大多需要相应的实验数据集，因此其可推广性较其他网络水军识别研究方法要低。

### 4．基于综合特征的网络水军识别研究

如上文所述的网络水军识别研究多基于特定类型的网络水军特征，但基于特定类型网络水军特征的识别方法无法全面分析网络水军行为，因此其识别准确率具有瓶颈。在此基础上，综合多种特定类型的网络水军识别方法对于各个目标领域的网络水军都具有较高的识别准确率。

### 5．各目标领域网络水军识别研究总结

邮件领域的网络水军识别研究在传统的基于内容特征的网络水军识别方法中引入用户行为、关系和环境特征，提高了网络水军识别表现。由于邮件领域网络

水军表现出高度的行为相似性，因此，基于行为特征的识别方法能够较好地发现邮件网络水军，在邮件领域各类网络水军识别方法中平均表现最好。

电子商务领域网络水军识别研究具有极高的应用价值，一直是近几年研究的热点。按照其识别体系不同，又可将电子商务网络水军识别研究分为监督识别和非监督识别。电子商务领域网络水军识别方法主要为基于网络水军行为和关系特征的方法。网络水军目标的相同，使其行为和关系表现出高度可识别的特征。因此，基于综合特征的网路水军识别方法能够达到较高的准确率和召回率，相比基于特定类型特征的网络水军识别方法表现得更好。

社交网络领域网络水军识别研究难度大，识别周期长，结果不易评价，但结合内容、行为和关系等特征的综合特征网络水军识别方法能够提高网络水军的识别表现，对复杂社交网络中的网络水军识别表现最好。

综上所述，传统网络水军集中于邮件领域并逐渐向电子商务和社交网络转移。这 3 个目标领域中的网络水军行为较为严重，对网络环境影响也较为严重。在 3 个目标领域中，网络水军识别都主要依据网络水军的自身特征，如内容、行为及关系等特征。与传统的基于内容特征的网络水军识别方法相比，当前网络水军识别研究的关注点转向网络水军的目的分析，实现从源头防治网络水军泛滥。其中，社交网络中的水军最为严重，且其行为较其他目标领域都更为复杂。基于特定类型特征的网络水军识别方法在邮件领域和电子商务领域可一定程度地发现网络水军，但在复杂的社交网络中，以上方法的网络水军识别表现较差。因此，社交网络水军识别研究基于综合特征进行识别，并不断适应性地增加网络水军识别新特征，以提高其识别准确率。

## 6.6 本章小结

随着 Web 2.0 的发展，人们越来越喜欢在网站发表自己对商品、对品牌的看

法，更喜欢在消费前参考别人的看法。网络水军通过在商品评论区发表对对手品牌商品的不实言论，通过污蔑、打压、谩骂的方式影响其网络口碑，给商家的正常经营造成了损失。

本章介绍了品牌推荐与自动保护系统的功能、设计和实现，最后通过实例测试了系统的品牌保护效果。

# 参考文献

[1] 常亚平，肖万福，覃伍，等．网络环境下第三方评论对冲动购买意愿的影响机制：以产品类别和评论员级别为调节变量[J]．心理学报，2012，44（9）：1244-1260．

[2] 杨铭，祁巍，闫相斌，等．在线商品评论的效用分析研究[J]．管理科学学报，2012，5（15）：66-72．

[3] 龚诗阳，刘霞，赵平，等．线上消费者评论如何影响产品销量[J]．中国软科学，2013，6：172-180．

[4] 宋冰，郑寒月．电影水军暗战被揭电影产业岂容“水军”泛滥[EB/OL]. http://culture.people.com.cn/n/2012/1219/c87423-19941390.html，2012-12-19.

[5] 王奕．深扒网络营销中水军的前世今生[N]．电商报，2015-10-08．

[6] 李金兰．品牌推荐与自动保护技术研究．北京邮电大学硕士毕业论文，2016．

[7] Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting spammers on Twitter. In: Proc. of the 7th Annual Collaboration Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS 2010), 2010, (6): 12-20. http://ceas.cc/2010/.

[8] 郝媛媛，叶强，李一军．基于影评数据的在线评论有用性影响因素研究[J]．管理科学学报，2010，13（8）：79-87．

[9] 姜巍，张莉，戴翼，等．面向用户需求获取的在线评论有用性分析[J]．计算机学报，2013，36（1）：120-129．

[10] 殷国鹏．消费者认为怎样的在线评论更有用[J]．管理世界，2012，12：116-122．

[11] Wang G, Xie S, Liu B, Yu PS. Review graph based online store review spammer detection. In: Cook D, Pei J, Wang W, Zaiane O, Wu X, eds. Proc. of the 11th Int'l Conf. on Data Mining (ICDM 2011). Washington: IEEE Computer Society, 2011: 1242-1247. [doi: 10.1109/ICDM.2011.124].

[12] Hayati P, Chai K, Potdar V, Talevski A. Behaviour-Based Web spambot detection by utilising action time and action frequency. In: Taniar D, Gervasi O, Murgante B, Pardede E, Apduhan BO, eds. Proc. of the Computational Science and Its Applications (ICCSA 2010). Heidelberg: Springer-Verlag, 2010: 351-360. [doi: 10.1007/978-3-642-12165-4_28].

[13] Bouguessa M. An unsupervised approach for identifying spammers in social networks. In: Proc. of the 23rd IEEE Int'l Conf. on Tools with Artificial Intelligence (ICTAI 2011). Washington: IEEE Computer Society, 2011: 832-840. [doi: 10.1109/ICTAI. 2011.130].

[14] Liu QW. Web spammers' detection and prevention. Press Outpost, 2012,6:021 (in Chinese with English abstract).

[15] Halpin H, Blanco R. Machine-Learning for spammer detection in crowd-sourcing. Human Computation AAAI Technical Report, WS-12-08, Menlo Park: AAAI Press, 2012: 85-86.

[16] Allan Heydon, Marc Najork. Mercator: A scalable, extensible Web crawler[J]. World Wide Web:1999, 4(2): 219-229.

[17] wawlian. 网络爬虫基本原理[EB/OL]. http://www.cnblogs.com/wawlian/archive/2012/06/18/2553061.html.

[18] 刘金红，陆余良．主题网络爬虫研究综述[J]．计算机应用研究，2007，24（10）：26-30．

[19] 王洪威．主题网络爬虫的分析与设计[D]．北京：北京邮电大学，2012：21-23．

[20] 岑咏华．科技信息门户网站的技术研究[D]．南京理工大学，2003：23-29．

[21] 任晓娜．基于 Lucene 的全文搜索引擎的研究与实现[J]．湖南广播电视大学学报，2010，30（5）：158-159．

[22] 董娟．基于页面结构分析的网页信息抽取方法研究[D]．中国石油大学，2010．

[23] Raykar VC, Yu S. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. The Journal of Machine Learning Research, 2012,13:491-518.

[24] Langbehn HR, Ricci S, Gonçalves MA, Almeida JM, Pappa GL, Benevenuto F. A multi-view approach for detecting non- cooperative users in online video sharing systems. Journal of Information and Data Management, 2010,1(3):313-328.

[25] Russell MA. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. 2nd ed., Sebastopol: O'Reilly Media, 2013: 5-43.

[26] Joseph Warren, Darrell Brunsch, Christopher Knestricck, et al. Messaging Over HTTP Protocol For Data Exchange[P].US Patent: US2014/048757, 2015-02-05.

[27] Jiang F, Du JW, Sui YF, Cao CG. Outlier detection based on boundary and distance. Acta Electronica Sinica, 2010,38(3):700-705 (in Chinese with English abstract).

[28] Benevenuto F, Rodrigues T, Almeida V, Almeida J, Zhang C, Ross K. Identifying video spammers in online social networks. In: Castillo C, Chellapilla K, Fetterly D, eds. Proc. of the 4th Int'l Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2008). New York: ACM Press, 2008: 45-52. [doi: 10.1145/1451983.1451996].

[29] Song J, Lee S, Kim J. Spam filtering in Twitter using sender-receiver relationship. In: Sommer R, Balzarotti D, Maier G, eds. Proc. of the 14th Int'l Symp. on Recent Advances in Intrusion Detection (RAID 2011). Heidelberg: Springer- Verlag, 2011: 301-317. [doi: 10.1007/978-3-642-23644-0_16].

[30] Murmann AJ. Enhancing spammer detection in online social networks with trust-based metrics [MS. Thesis]. San Jose: San Jose State University, 2009.

[31] Hayati P, Chai K, Potdar V, Talevski A. HoneySpam 2.0: Profiling Web spambot behaviour. In: Proc. of the Principles of Practice in Multi-Agent Systems. Heidelberg: Springer-Verlag, 2009: 335-344. [doi: 10.1007/978-3-642-11161-7_23].

[32] Wang JH. Social Network Analysis for E-mail Filtering. 2006: 69-78. http://www.im.cpu.edu.tw/cyber06/cyber06-a6.pdf.

[33] Chang M, Yih W, Meek C. Partitioned logistic regression for spam filtering. In: Li Y, Liu B, Sarawagi S, eds. Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Minging (KDD 2008). New York: ACM Press, 2008: 97-105. [doi: 10.1145/1401890.1401907].

[34] 张黎，牛耘．基于迭代法的产品评论情感倾向分析[EB/OL]．中国科技论文在线．

[35] Li Shoushan, Huang Churen, Zhou Guodong, et al. Employing Personal/ Impersonal Views in Supervised and Semi-supervised Sentiment Classification [C]. USA: Association for Computational Linguistics Stroudsburg, PA, USA ©2010, 2010:414-423.

[36] Balaguer EV, Rosso P. Detection of near-duplicate user generated contents: The SMS spam collection. In: Proc. of the 3rd Int'l Workshop on Search and Mining User-Generated Contents (SMUC 2011). New York: ACM Press, 2011: 27-34. [doi: 10.1145/ 2065023.2065031].

[37] Sathawane KS, Tuteja RR. A robust spam detection system using a collaborative approach with an E-mail abstraction scheme and spam tree data structure. Int'l Journal of Computer Science and Applications, 2013,6(2):293-298.

[38] Tseng CY, Sung PC, Chen MS. Cosdes: A collaborative spam detection system with a novel e-mail abstraction scheme. IEEE Trans. on Knowledge and Data Engineering, 2011,23(5):669-682. [doi: 10.1109/TKDE.2010.147].

[39] Maan G, Tak G. Enhanced discussion on different techniques of spam detection. Int'l Journal of Computer and Technology, 2013, 4(2):248-253.

[40] Hayati P, Potdar V. Spam 2.0 state of the art. Int'l Journal of Digital Crime and Forensics, 2012,4(1):17-36.

[41] 莫倩，杨珂．网络水军识别研究[J]．软件学报，2014，7（25）：1505-1526．

# 第 7 章　网站验证码识别

## 7.1　引言

验证码技术作为网络安全中最常见的基本手段之一，得到了深入的研究和广泛的应用。在网络舆情监测系统和品牌推荐与保护系统的信息爬取、自动登录和自动评论等过程中经常会遇到验证码识别的问题，需要程序自动化地完成。验证码识别的技术有很多，本章主要介绍基于模板匹配方式的验证码识别技术。本章详细介绍了验证码识别中各个阶段的技术方法和理论，主要包括图片预处理阶段、图像字符分割阶段和字符的匹配识别阶段。针对各阶段的需求，综合设计实现了完整的验证码识别功能[1]。

首先，介绍了图片预处理中需要使用的数字图像处理技术，重点分析了图像灰度化、二值化处理、图像噪声消除及验证码中的图像干扰线去除等过程的理论和常用的技术手段。针对曲线型干扰线问题，设计了一种相应的解决方法，并获得了较好的效果。

其次，介绍了图片字符分割的多种方法和理论，并针对每种算法的设计实现步骤，分析各种方法的优缺点和适用场景。对连通区域检测算法提出了改进方法，能够适应轻度粘连的字符分割场景。同时在结合连通区检测和投影方法优势的基础上，通过加入过度分割的后处理策略，设计了一种基于连通区检测和投影的分割方式，对于字符预处理后存在像素损失的场景有较好的适应能力。

再次，分析和研究了字符特征建模的三种方法和理论，阐述了各自的适应场景和优势，提出了基于边缘信息的轮廓走势特征，对字符的旋转和形变有良好的适应性。

最后，实验分析证明了上面所述识别策略和方法的有效性，以及特征模型的字符描述能力。

## 7.2　验证码识别

### 7.2.1　验证码的概念

验证码是一类用来判别使用者或服务对象是人还是机器程序的公共自动识别程序或代码。验证码工作的流程是：服务器端发送给用户一些需要用户回答的问题，这些问题可以是文字的、图像的或音频的，用户将答案反馈给服务器，服务端根据用户的回答判断用户是人还是自动程序，判断标准一般是用户的答案和标准答案匹配则判别用户为人。从流程中可以看出，服务器端的问题设计是验证码

工作的关键，这些问题要保证能够很容易地被人回答，但是自动程序不能正确回答或者给出正确答案非常困难。

为了提高验证码的防破解能力，保障其安全性，验证码的研究者和设计人员提出了很多反自动识别的技术。Baird H S 和 Riopka T P[6]等通过将字符按垂直、水平方向切成片段的方法提高了反字符分割的能力。其他反分割手段，如字符粘连，在设计验证码的过程中也被研究人员分析和广泛使用。

## 7.2.2 验证码分类

验证码有非常多的种类和形式，根据它们的适用场景和自身特点，将其大致归类为：基于文字的图片形式验证码、基于内容判别形式的图像验证码、基于声音识别形式的验证码，以及关于数学运算类的验证码。

我们主要关注基于文字的图片形式验证码，这也是国内外网站中应用最为频繁的验证码。将文字字符以图片的形式展示，识别图片中的文字，从而完成验证。其特点是验证码生成简单，有标准答案，用户操作简便，且对用户的知识背景要求不高。同时，通过反分割、反识别技术使得验证码安全性较高，破解的复杂度非常大。这类验证码在用户加载页面的时候随机生成一幅图像，包含了字符信息，并通过加入背景色、干扰点、干扰线，以及对字符旋转、变形、伸缩变换等方法，增加图像自动识别的难度。百度贴吧（tieba.baidu.com）的注册系统使用的验证码即为此类验证码，如图 7-1 所示。本文中分析与研究的验证码系统类型或对象即为此类验证码。

图 7-1　百度贴吧注册验证码

### 7.2.3　验证码识别框架

针对不同的验证码系统，需要有针对性的分析和设计。验证码识别中涉及的技术主要包括图像处理技术、模式识别和机器学习技术等，识别的主要过程包括验证码的图片资源收集、验证码的图片预处理（包括图片的灰度化处理、二值化处理、干扰点及干扰线的去除）、验证码字符分割和验证码字符识别 4 个方面，如图 7-2 所示，后续将详细介绍。

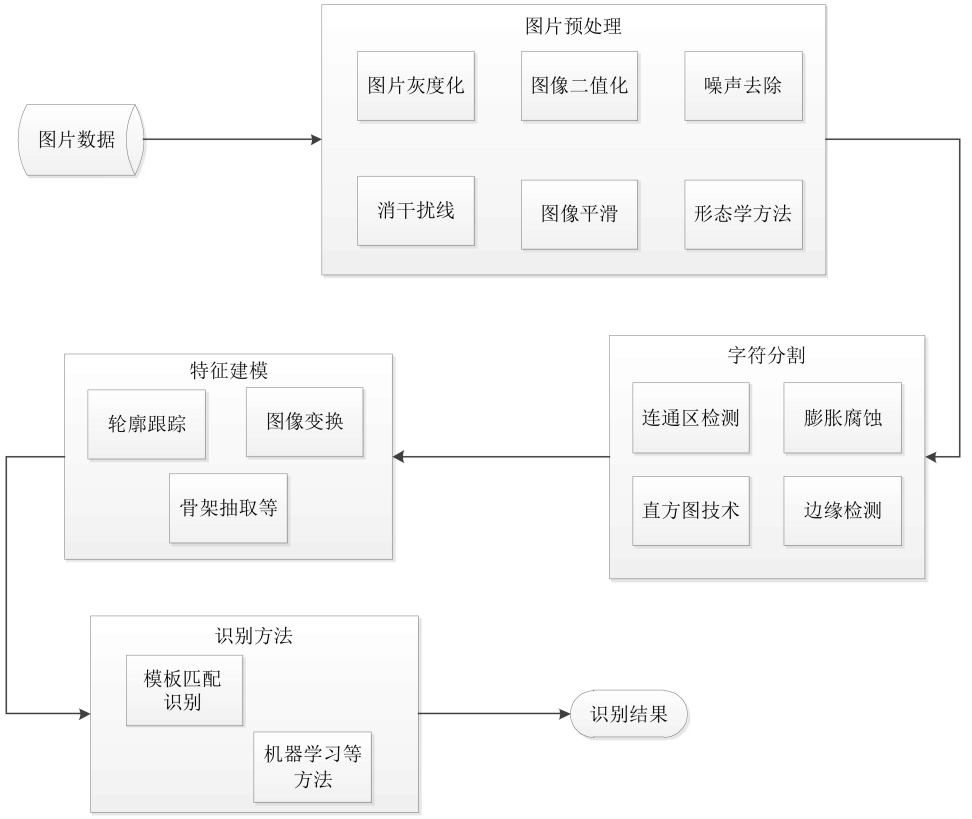



图 7-2　验证码识别框架图

## 7.3 图片预处理

验证码系统设计的目标就是反识别，防止自动程序破解识别验证码，提高系统的安全性。因此，在验证码的设计中使用了许多反识别的手段，其中主要的手段和技术包括如下几种。

（1）增加背景色，减小目标字符与背景差色，从而降低自动程序对目标字符与背景色的区分度，增加自动程序提取目标字符的难度。如图 7-3 所示，背景色和目标为同种颜色。

图 7-3　颜色单一的验证码

（2）增加噪声点，混淆目标字符。噪声点具有与目标字符相似的像素值，这种设计使得自动程序判别目标字符的像素的难度增大。如图 7-4 所示，噪声点和目标字符的像素值相同，分布在字符周边及字符上，极大地提高了自动程序准确提取目标字符的难度。


图 7-4　噪声验证码

（3）增加干扰线，字符之间存在干扰线粘连，同时干扰线与字符的曲线相近，弥补了噪声点面积小带来的缺陷，使得字符的分割及获取目标字符像素点的难度增大。如图 7-5 所示，字符 m、F 之间由干扰线连接在一起，同时，干扰线的宽

度和字符曲线相近，普通的投影分割不能分割字符。另外，字符 Z 和与之相连的干扰线像素值一致，自动程序不能直观地获取字符的像素。干扰线在很大程度上提高了验证码的反分割和反识别能力。

图 7-5　干扰线验证码

（4）字符粘连，即相邻目标字符粘连在一起。字符粘连使得字符不能无损地分割开，自动程序判定字符之间的分割边界难度极大。这种手段是验证码反分割的有效方法。如图 7-6 所示，字符 8 之间粘连在一起，自动程序难以将 8 分割开，提高了自动识别的难度，同时易于用户识别。

图 7-6　字符粘连验证码

（5）字符旋转和变形，即对单个字符进行旋转和拉伸等变换后嵌入验证码图片中。这种技术手段很好地防止了通过简单的像素匹配的方式破解验证码。对于不旋转和变形的验证码，只需对字符建立简单的像素模型，通过像素点匹配即可识别。通过字符旋转和变形，使得验证码的安全性得到极大提升，同时也使得其他技术手段得以发挥效果。如图 7-7 所示，字符 9 有变形，图中 3 个字符 9 均不一样，同时字符 3、4、8、9 等均有旋转操作。

图 7-7　旋转与变形验证码

在图片预处理阶段，需要提取出较为完整的目标字符，去除和目标字符无关的像素点，包括背景颜色和纹理、噪声干扰像素点、干扰线等。同时，需要设计合理的存储方式，降低识别的复杂度，并易于操作和处理。

### 7.3.1 图像灰度化

验证码图片一般是彩色图片，包括 R、G、B 三种颜色分量，有的图片还有 Alpha 通道。彩色图包含非常丰富的信息，在处理彩色图时，由于计算的空间非常大，导致计算量非常大。同时，人对于 R、G、B 三种颜色的感知是不一样的，并不能对细微的颜色变化有感知。因此，在处理和计算图像数据时，为了降低计算的时间开销和空间复杂度，一般将原始图像数据转换为灰度像素图像数据来处理和分析。

灰度化的处理是将 RGB 三色空间转换为 1 字节存储的图像，这将极大地减少计算量，同时灰度化的图像依旧保留了原有的亮度和色差信息。灰度化的方法有很多种，主要思路均是将 3 字节图像压缩为 1 字节图像。常见方法有 4 种。

（1）取一个分量作为灰度值。该方法根据某一通道的值确定灰度图像，在应用中根据实际场景确定各分道的重要性，再选择通道值。在验证码处理中，可以用来消除特定颜色的干扰线和噪声点。

（2）分量最大值方法。该方法在计算灰度像素值时，选择三色通道 3 个值中的最大值作为当前像素点的值。

（3）取平均值的方法。该方法将三色通道的颜色值作为同等重要的地位，取其均值作为灰度值。该方法将颜色平等对待，具有颜色平滑的作用。

（4）取颜色亮度值作为灰度像素值。YUV 颜色空间是欧洲国家电视系统中常使用的一种颜色编码方式，侧重于颜色模型中的亮度值，与人眼对颜色的敏感程度相似，其中 Y 表示亮度，即灰度值，U 和 V 是两个色差值。通过将 RGB 表示

的颜色空间变换到 YUV 表示的颜色空间中，从而提取出颜色模型的灰度值，这种灰度化的方式更符合人眼的习惯。

## 7.3.2　图像二值化

### 1. 基本介绍

图像二值化是进一步利用图片处理技术去除无关信息的过程，其作用是缩小目标信息的范围，减少冗余信息，降低后续处理的复杂度。

图像二值化的效果是将图片中的像素信息分为两类：一类是目标像素点，以数值 255 表示；另一类是背景像素点，以数值 0 表示。二值化的效果如图 7-8 所示，目标为白色，像素的亮度较高；背景像素点为黑色，具有较低的亮度。

图 7-8　图像二值化处理

二值化是图像数据处理的基本技术，其处理方法是选取某一灰度像素值 Thred，若像素点的颜色灰度值大于 Thred，则设置当前像素点的值为 255；否则，设置当前像素点的值为 0。图像阈值处理分为图像像素全局阈值和像素数据局部阈值[21]。全局阈值的思路是，针对图像的所有像素点，均按照同一个阈值参数对像素点进行处理；局部阈值的思路是，对于像素点，计算局部阈值 $T_i$，使用 $T_i$ 进行阈值处理。局部阈值处理对于背景色有渐变的情况效果较好，适应了局部特征，但增大了计算量。

二值化图像处理的主要技术点是如何选取某一灰度值 Thred 作为阈值参数，很多学者对这一问题进行了研究。2002 年，李建华、马小妹等[22]针对指纹图像数据的处理技术研究提出了基于方向图的动态阈值二值化方法，利用指纹形式的图

像中的方向和灰度变化的特点，直接从指纹源图像中计算出二值化图像。孙少林、马志强等[23]讨论并比对了应用较为常见的几类图像二值化处理算法。Rivera M, Mayorga P P[38]实现了一种马尔科夫随机场模型进行二值化处理。在图像数据处理的技术和手段中，常用的二值化技术包括普通阈值、最大类间方差法（OTSU）阈值、直方图阈值、Bernsen 阈值。

我们研究的验证码图片背景色具有反识别的作用，同时验证码具有有利于人眼识别的目标，图片背景色与目标字符之间具有可区分度，但直接的阈值处理效果不佳。我们采取的是最大类间隔阈值的方法，即 OTSU 算法，该算法能动态自适应地计算灰度阈值，能基本把与字符无关的背景色去除。

### 7.3.3 图像去噪

图像去噪是图像处理的基本技术之一，冈萨雷斯在其《数字图像处理》一书中介绍了图像噪声消除的方法[16]，包括空间滤波、中值滤波、均值滤波、频域滤波等，同时还介绍了形态学开运算去噪声的原理，这些方法均适用于灰度图像处理。

验证码噪声与一般的图像噪声有区别。一般的图像噪声是图像在拍摄或形成时，受仪器或环境的影响所形成的；验证码噪声一般是人为根据算法或随机生成的，其目的在于防止自动程序识别。验证码噪声多与一般图像噪声的椒盐噪声相似，如图 7-9 所示。


图 7-9　噪声验证码

验证码的噪声点的灰度值一般与目标字符一致或相似，在设计时，考虑到增

大目标像素提取的难度，会将噪声点处的灰度像素值设计为与目标字符的灰度像素值一致。验证码噪声去除的目标是消除噪声点的多余信息，保存完整的目标字符信息。这一步降低了图像的信息量，使得现存的信息进一步缩小，越来越接近目标字符所表示的信息量。

验证码的噪声去除方法有两类：第一类是针对灰度图像的噪声去除方法，这类方法也叫图像平滑方法，其目的是通过图像数据的平滑处理操作，降低噪声点处的灰度值，从而在图像的二值化处理时去除噪声点；第二类是针对二值图的噪声点消除算法，这类方法主要利用噪声点所在像素点连通区面积较小及噪声所在像素块背景像素点多的特点，直接将噪声点处的颜色值设置为背景颜色从而去除。常用的消去噪声的方法有中值滤波、均值滤波、高斯平滑、频域滤波、二值图孤点消除、二值图连通域阈值、二值图开运算去噪等。

## 7.3.4　干扰线去除

验证码图片中常采用干扰线的形式防止自动程序检测目标字符的笔画线条。干扰线有直线和曲线两种形式，如图 7-10 所示。干扰线的目标是干扰字符的正常像素，防止自动程序检测；可以连接字符像素，防止自动程序对字符进行分割处理。同时，干扰线设计为保证人的体验度，不能与字符笔画太相似，需要有区分度。

（a）直线干扰线　　（b）曲线干扰线

图 7-10　干扰线的形式

干扰线连接多个字符是粘连字符验证码的一种形式，这种形式的好处在于能起到反分割的作用，同时对于人识别单个字符的影响较小，是字符粘连常见的一

种形式。直线干扰线去除可以采用图像直线检测技术，如采用 Hough 变换形式的检测方法；曲线干扰线暂无固定的去除方法。本文将讨论直线干扰线去除和曲线干扰线去除两种情况。

在消除直线干扰线时，可以使用 Hough 变换检测直线，它将直线关于位置的线性函数关系映射到参数空间，然后采用统计的方法选择参数，从而确定直线上的像素点，根据像素点数量判断是否为干扰线，特点是直线干扰线包含的像素点数量较多。验证码的图片直线有很多颜色方面的特点，本文利用这些特点设计了直线干扰线去除的方法。在如图 7-10（a）所示的验证码中，可以采用按颜色聚类的方式，即通道提取。通过分析，在 Alpha 通道中，字符的像素灰度数值明显大于干扰线的灰度像素值。通过提取 Alpha 通道的灰度值，然后使用阈值二值化技术即可提取出字符的像素。如图 7-11 所示是处理的过程，阈值为 200，效果明显。

（a）原图　　（b）提取 Alpha 通道　　（c）二值化处理后的图

图 7-11　直线干扰线处理过程

消除曲线干扰线没有固定的方式，一般的做法是分析干扰线的特征，然后利用特征判断干扰线，消除干扰像素点。本文处理的曲线干扰线如图 7-10（b）所示，干扰线和字符粘连，而且可能连接多个字符。可以分析其特点：曲线与字符的笔画有明显的区别，这种设计是为保证人的识别率；同时干扰线一般是验证码生成程序按算法生成的，与自然的字符有很大的差别。图中的曲线与字符曲线相比，宽度较为固定，且较为平滑，本文利用这一特点设计了干扰线去除算法。算法的基本思想是根据干扰线的曲度较小的特征，检测曲线线条，根据宽度判断是否为干扰线。具体的算法步骤如下：

（1）遍历图像数据中的每一个待检测像素点，以该点为起始点遍历 8 个方向检测曲线。

（2）在原方向的 45° 夹角范围内搜索邻居点，若存在目标像素点，则加入线条集合中，并更新方向，重置为新的位置点。

（3）根据搜索得到的曲线（单像素点宽度）将线条扩充，即加入线条的邻居像素点，根据线条的宽度判断是否为干扰线，是则去除。

根据以上算法设计的流程处理验证码图片后的结果如图 7-12 所示。

（a）原图　　（b）去干扰线后的图

图 7-12　曲线干扰线处理

该算法很好地去除了干扰线，较为完整地保留了字符的像素信息，但是依旧有部分像素可能因去除干扰线而被“误伤”。该算法在本文研究的验证码识别中取得了较好的效果，算法使用了干扰线特有的特征，因此对于研究不同的验证码识别技术，需要分析提取特定的干扰线特征。

## 7.4　字符分割

### 7.4.1　字符分割简介

图像字符分割的过程是将验证码图片信息中各个字符所在的图像区域分割出来，以实现对单个图像字符的图片处理操作，Serge Belongie、Greg Mori 等[2]在对图片分割的基础上提出并设计了字符形状上下文的特征表述形式的识别方法。图像字符分割的技术广泛应用于车牌号码识别领域，张云刚和张长水[3]利用 Hough 变换算法识别车牌号码。陈寅鹏和丁晓青[4]提出并设计了一种使用模板匹配方法

的车牌字符分割的算法，并实现了一个完整的车牌识别系统。

图像字符分割处理是验证码识别过程中一个非常重要的环节，它加工和操作的图片对象是经过图片预处理后的二值化的验证码图片，将原始的图片分割为只包含单个字符的图片以便于分析和获取字符图片的特征，以及对字符建模。同时，分割图片字符后便于训练字符图片的模型，设计识别算法。

字符分割是图像分割的一种具体应用，同时具有验证码的特点，可以对图像分割的技术加以改进，从而适应字符分割。较为常见的图片字符分割的方法有K-Means 聚类分割、图像投影分割算法、连通区分割方法、滴水分割算法等。这些方法都有适应的场景，在设计验证码分割算法时，在特定情形下能取得很好的分割效果。

## 7.4.2 K-Means 聚类分割

K-Means 聚类算法[5]是数据挖掘中常用的一种技术手段[24]，在验证码识别中可以根据二维空间的距离和验证码预先的字符数量对像素聚类，达到字符分割的目的。K-Means 聚类算法的主要思路是：先在图像数据中随机选择出 $K$ 个像素点作为初始的分类质心；然后遍历图像的每一个像素点，并计算到各质心的距离，距离计算时需要考虑像素点的连通性，将其分类到距离最近的质心；处理完所有像素后，重新计算生成各类的质心；不断迭代执行为各点分配类和计算质心的过程，直到类的质心收敛，聚类算法完成。

在实际处理验证码图片时，初始选择质心时，可根据验证码图片中字符水平排放的特点，在水平方向上按等间隔的方式选择像素点作为质心。衡量向量的距离时可以使用二维空间的距离，即欧式距离来度量。K-Means 只能处理图片中字符个数固定的验证码图片，对于字符个数会发生变化的验证码有一定的局限性，同时对于存在粘连的字符分割效果不好。

## 7.4.3　投影分割

投影分割是一种常见的字符分割方法，迟晓君和孟庆春[25]提出了使用垂直投影的方法对车牌进行字符分割，能快速地找到字符之间的最优分割点。在验证码识别过程中，使用的投影方法主要是垂直投影，它利用了验证码字符按水平方向排布的特点；同时对于部分验证码，字符之间的粘连度较低，在垂直投影的特征上有很好的区分度。垂直投影的数据是一种直方图数据，通过分析和计算直方图的极小值点，便能确定字符之间的分割点。

在验证码识别过程中，垂直投影分割算法处理的数据图像是预处理后的图片，其主要思路如下。

（1）二值图像在 $x$ 轴方向作垂直投影，形成直方图。具体做法是：对于每一个 $x$ 轴上的点，统计垂直方向上点的数量。在形成直方图时，以 $x$ 坐标的值作为横轴，以统计的像素点的数量作为纵轴。

（2）根据投影的直方图计算分割点。在直方图中，点的数量较多的地方一般是在字符上，较少的地方可以认为在字符之间。对于没有噪声点的图像，字符之间的点的数量为 0。

投影算法处理的结果如图 7-13 所示。可见，在无噪声且预处理结果较好的字符无粘连的图片上效果很好，将字符完整地分割开了。

（a）原图　　（b）预处理后的图　　（c）分割后的图

图 7-13　投影算法处理

预处理采用了灰度化、OTSU 阈值二值化及关于曲线干扰线的去除，预处理结果较好，字符较为完整地保留下来，干扰线去除了，且字符间不存在粘连。采

用垂直投影的方式，投影直方图如图 7-14 所示。由图 7-14 可知，字符之间的区分明显，可取投影后数值为 0 的位置作为分割点，取分割点 $x$=30、$x$=60、$x$=95 将图片分割为 4 张字符图片，如图 7-13（c）所示，分割效果很好。

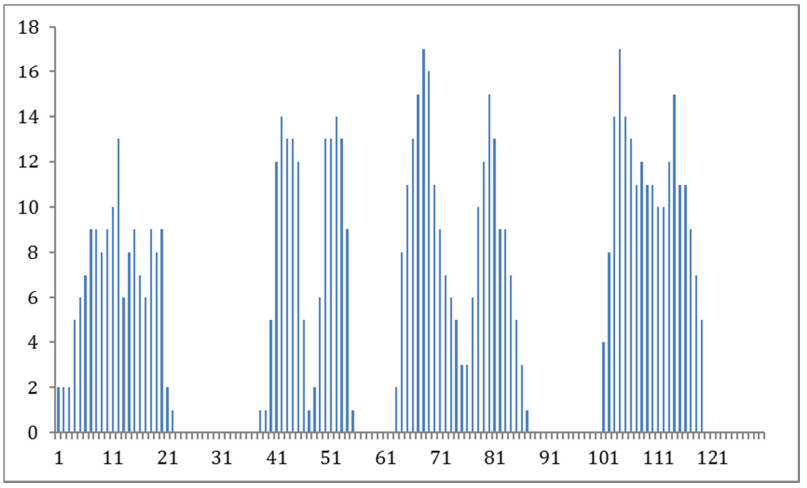



图 7-14　投影直方图

这种垂直投影的方式只考虑了字符在水平方向的区分度，对于字符在水平方向有重叠但字符本身不重叠、字符粘连及有噪声点去除效果不好等情况，采用投影的方式，分割效果较差。

### 7.4.4　改进的连通区检测

连通区域检测算法[26]是图像处理常用的技术手段。在预处理过程的噪声去除阶段，曾利用连通区域检测算法计算像素面积的方式去除面积小的噪声区域。在字符分割阶段，使用连通区域检测算法来发现字符像素，实现字符分割。

与连通域阈值算法的基本原理一致，为处理轻度的字符粘连，对于像素点加入当前连通区域作一定限制，采用的思路是：将像素点分为边缘像素点和连通区

内部像素点，根据像素点 8 邻域是否全部为目标点判别，若 8 个点全为目标点（值为 255），则该像素点为区域内部像素点；否则该点当作区域边缘的像素点。对于内部像素点，直接将其 8 邻域像素点加入当前连通域作为同一个字符的像素；对于外部点，在处理其邻域内的 8 个像素点时，如果与邻域像素点相邻的其他像素点数量少于 2 个，则将邻域像素点加入当前连通域，作为同一个字符的像素点；否则需要判断点与当前连通域的连接程度是否大于其他区域的连接度，若是则加入当前连通区，否则不加入当前连通区域。具体的算法设计思路如下：

（1）初始化设置字符区域的存储空间 $C$ 为空，$C \leftarrow \phi$。

（2）初始化设置活动字符的像素存储空间 $P$ 为空、活动队列 $Q$ 为空，$P \leftarrow \phi$、$Q \leftarrow \phi$。遍历图像中所有未标记的像素点，对于未被标记的像素点 $S$，作为当前字符的种子像素，以它为起点检测字符连通区域。将 $S$ 加入队列，$Q \leftarrow S$；设置像素点 $S$ 为已标记状态。如果不再存在未被标记的待处理像素点，则执行步骤（6）。

（3）测试队列。如果队列不为空，则弹出一个点 $T$，并将像素点 $T$ 加入活动字符存储空间 $P$，$Q \leftarrow T$。如果队列为空，则执行步骤（5）。

（4）检测像素点 $T$ 所有相邻的 8 个邻域像素点，若其邻域像素点的像素值均为 255，则当前处理的像素点为连通区域的内部点，直接将其所有的邻域像素点中未被标记的点 $x$ 加入当前队，$Q \leftarrow x$，并设置为已标记状态列。如果像素点 $T$ 的 8 邻域像素点中存在像素值为 0 的点，则当前处理的点 $T$ 为边缘点，对于其 8 个邻域像素点中未被标记的像素点 $y$，统计 $y$ 的 8 邻域像素点中未被标记的点数 $n_1$ 和已被标记的点数 $n_2$，若 $n_1 \leqslant n_2$，则表明 $y$ 点与当前字符的连接度更强，将 $y$ 点加入队列，$Q \leftarrow y$，并将 y 点设置为已标记状态；若 $n_1 > n_2$，则不处理 $y$ 点。处理完 $T$ 的所有 8 领域未被标记的像素点后，返回步骤（3），继续访问其他点。

（5）队列为空，当前连通区域不再有新的连接点，检测出一个字符连通区域，将字符加入字符存储空间 $C$，$C \leftarrow P$。返回步骤（2），重新检测其他字符连通区域。

（6）所有像素点均被检测，字符分割完成，字符存储空间 $C$ 中存储了所有分割后的字符像素，算法结束。

该算法是对普通连通区域的改进算法，主要特点是对于像素点的检测设计了连接度的计算，增强了该算法对于轻度粘连字符的分割力度。如图 7-15 所示为待分割的像素点图像示例，包含 8 个像素点，使用该算法会将像素点分割为两类 $\{a,b,c\}$、$\{d,e,f,g,h\}$。其中对于 $d$ 的划分，因为 $d$ 与第一类只有一个像素点相邻（$c$ 点），而与右边第二类有两个像素点相邻（$f$ 点和 $e$ 点），所以第一类和第二类在 $d$ 点处分割，且 $d$ 点输入第二类。

| | | | | | |
|---|---|---|---|---|---|
| | *a* | *b* | | *g* | |
| | | *c* | | *e* | *h* |
| | | | *d* | *f* | |
| | | | | | |

图 7-15　分割原图

## 7.4.5　滴水分割算法

1995 年，Congedo G、Dimauro G 等[27]在分割手写数字字符时提出了滴水分割算法（Drop Falling），在字符分割中有着广泛的应用。该算法主要用来分割存在粘连的字符，将输入的图片像素分割为两部分。常丹华、何耘娴等[28]对该算法提出改进后应用于中英文混排文档图像中粘连字符的分割，针对滴水算法的两个不足之处，一是起始像素点的选择易受局部干扰，二是垂直滴水分割损害笔画像素，提出了先使用贝叶斯分类器判定字符图像的类型，再使用阈值方法

提取出粘连字符，而后通过字符轮廓求极值确定分割点，最后使用中心线改进滴水分割算法。李兴国和高炜[29]将滴水分割算法应用到了验证码识别的字符图像分割中。

滴水分割算法的主要思想是：从图片的上方选取分割的起始点，然后以此点为始逐步根据策略选择下一步分割点。策略的主要分割方向是向下进行的，通过判断当前像素点的左、左下、下、右下、右 5 个点的像素值确定下一步选择的像素点。

滴水算法关于方向的选择如图 7-16 所示。其中，黑色方框表示当前处理像素点，箭头表示下一步处理方向，*w* 表示白色像素点（255），*b* 表示黑色像素点（0），*表示白色和黑色像素点均可。

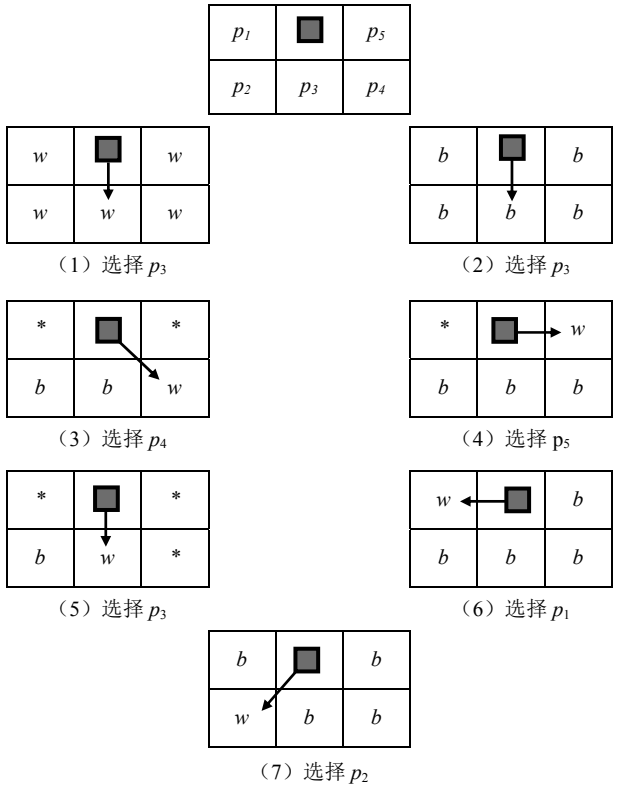



图 7-16　滴水算法方向选择

在验证码字符分割中使用滴水分割算法，根据验证码的先验知识，如字符的宽度范围、图片的宽度等信息辅助确定分割算法的初始分割点。使用先验知识可以较为有效地解决初始点选择不合理导致将字符分割为两部分的缺点。

## 7.4.6 基于连通区检测和投影算法结合的分割方法

在验证码识别过程中，单一的分割算法适用的场景较为固定，连通区域检测算法适用的场景主要是字符不粘连或者粘连度很小的情况，而投影算法适用于字符图像在水平方向可被分割的场景。本文将二者结合设计了一种分割方法，以及分割的后处理方法，能够对部分粘连度较大、预处理导致字符被分开的情况的验证码有较好的分割效果。如图 7-17 所示，原图经过图像预处理后，字符 Z 因去除干扰线导致字符不连续，直接使用 7.4.4 节改进的连通区检测算法会将字符 Z 分割为三部分，而通过结合投影算法，正确地将字符 Z 分割为一个整体。

（a）原图

（b）预处理后的图

（c）分割后的图

图 7-17　分割算法处理

对于字符之间粘连度较大的情况，如图 7-18 所示，对原图进行预处理后，W 和 D 的粘连度大于两个像素；使用 7.4.4 节改进的连通区检测算法会将 W 和 D 的像素分割为一个连通区，作为一个字符；而通过结合投影算法，正确地将 W 和 D 字符分割开。

（a）原图

（b）预处理后的图

（c）分割后的图

图 7-18　粘连度大的图片分割

算法的主体使用的是 7.4.4 节改进的连通区域检测分割算法，再通过投影分割算法对字符粘连度大的予以分割，最后设计了图像过度分割处理后的连通区域合并算法。算法的主要思想如下。

（1）使用 8 连通区域检测算法计算连通区域，并利用阈值的方式将像素点面积极少的消去，检测出的连通区域作为初步划分区域。

（2）利用先验知识，如字符个数和字符宽度的范围，使用阈值的方式检测需分割的连通区域，并采用垂直的投影算法与标记算法结合的方式进行分割。具体的思想是：

① 利用投影的思想在区域的中间部分计算最好的分割点，并将区域分割为两部分。

② 分割时采用基于普通连通区检测的标记算法，首先利用连通标记为分割线左右连接的区域标记类别，然后为处于分割线上的争议点划定标记，最后对于未标记的区域利用连通区生长算法将距离最近的标记作为其分类。

③ 经过前两步将区域划分为两部分，形成两个字符区域。

（3）连通区合并检测。对于过度分割的图片，需要将连通区进行合并，图 7-17 所示的字符分割中字符 z 的合并在此步骤完成。合并的主要思路是：相邻连通区且有水平交叉的优先合并，使用先验字符个数阈值对字符进行合并处理。

该算法是针对本文所研究的验证码特点所提出的，能很好地解决本文的验证码分割。该算法主要利用了字符的水平排布和字符连通性的特点来设计，在算法中使用最多的是连通区检测算法，同时利用了投影能根据水平粘连度检测分割点的优点，最后加入了后处理模块。

## 7.5 字符识别

字符识别是将图片字符转换为文本字符的过程，它是模式识别的一种具体应用，用来提取图片中的模式并描述。字符识别技术中较为常见的有模板匹配方式、基于神经网络的识别方式、利用支持向量机理论的识别等方式，图片字符的识别是数字图像处理技术与机器学习等方法的综合运用。马俊莉、莫玉龙等[30]利用模板匹配的方法设计了车牌识别系统。王敏、黄心汉等[31]在 2001 年利用神经网络设计了车牌识别算法，识别率超过 90%。Bhowmik T K、Ghanty P 等[32]利用支持向量机（SVM）的理论和方法设计了识别手写体字符的算法。

本文主要分析并研究通过模板匹配的方式识别图片验证码系统，识别方法的主要步骤有：图像字符像素特征建模，获取字符图片的特征，用于表征字符；特征库构建和制作，主要是制作模板匹配库；匹配识别过程，即对于线上的图片字符使用模板库中的特征进行匹配识别。

### 7.5.1 字符特征建模

字符特征模型是对字符图片的描述，在计算空间中将用字符特征代替字符图片参与运算，字符模型设计的好坏直接决定着识别的准确度。设计字符特征模型的目标：一是能具有代表性地描述字符图片，它表示特征模型描述字符的准确性；二是特征模型能使得不同字符之间有较大的区分度，同一个字符特征相同或相近，即将字符图片映射到特征模型空间后仍然保持原空间里的可区分性。以上两个目标中，可区分性将使用特征间的距离来度量，准确性将通过字符识别的准确率来间接衡量。字符特征模型形式根据验证码特点的不同而多种多样，如形状上下文特征[33]、字符小波特征[34]等。本文根据所研究的验证码特征选取或设计了三种特征：字符区域像素统计特征描述、边缘轮廓走势特性描述和像素位图特征。这三

种特征有各自的适用场景，需要根据具体验证码特点选择实现的方式。

### 1．字符区域像素统计特征

区域像素统计特征是将字符图片分割为固定的区域，一般是 3×3 宫格形式，然后统计每个区域内像素值为 255 的像素点数，最后计算每个区域像素点数占全部像素点数的比率，计算的 9 个比率值组成了一个 9 维的特征向量，这个向量即表征图片字符的区域像素统计特征。这种特征计算简单且存储量小，是图像与字符识别常用的简单特征。该特征模型能基本描述字符图片的形状特征。区域像素统计特征的算法设计思路如下（采用 3×3 宫格的区域）。

（1）将字符图片划分为 3 行 3 列的形式，一共对应 9 个小格子 $B_i$，$i$ 的取值范围为 1～9。

（2）遍历图像中的所有像素点，统计每个格子 $B_i$ 中包含的字符像素点个数。

（3）计算每个格子中的像素数占像素总数的比率 $R_i$，公式如下：

$$R_i=B_i/N$$

其中，$N=\sum_{i=1}^{9}B_i$ 。

字符的特征 $F=(R_1,R_2\cdots,R_9)$。

字符区域像素统计特征适用于字符能够分割的场景，该特征方法能较好地描述字符。同时，区域像素统计特征计算的是像素比率，不需要进行字符图片大小归一化处理。其缺点是不能适应字符的旋转和扭曲，如果有旋转和扭曲，则将导致特征库增大。

### 2．字符边缘轮廓走势特征

字符边缘轮廓是指字符像素点的连通区域的最外层像素点。如图 7-19 所示，“+”、“-”构成连通区域，“+”是最外层像素点，组成边缘轮廓。

| | | | | |
|---|---|---|---|---|
| | | | | |
| | | + | + | |
| | + | - | + | |
| | + | - | + | |
| | | + | | |

图 7-19　边缘轮廓图

字符边缘轮廓信息描述了字符的形状特点。因为字符是由笔画组成的，轮廓与笔画相近，所以使用边缘轮廓信息能较为准确地描述字符图片。在图像处理中，轮廓信息是一种描述图像的重要特征[35]，其主要的表示方法是边界链码[61,62]。边界链码将字符的轮廓像素点的相对位置关系用数字刻画，如图 7-20 所示。对于图 7-19 所示的边缘轮廓，如果以左上方的“+”作为起始点，则其轮廓链码可表示为：6542117。

| | | |
|---|---|---|
| 8 | 1 | 2 |
| 7 | + | 3 |
| 6 | 5 | 4 |

图 7-20　相对位置关系

边界链码的表示方法将轮廓信息描述为有序的数组，方便计算处理，是一种描述图像轮廓信息的很好的方式，但是这种表示方法不能适应字符的旋转。如将图 7-19 所示的图像向右旋转 90°（如图 7-21 所示），则边界链码变为：5778234。旋转后的边界链码与之前的不一致，因此，在处理验证码字符轮廓时，使用边界链码会造成相同字符的边界链码差异很大，这与特征设计的第二个目标（相同字符的特征相同或相近）冲突，不能达到准确描述图片字符的目的。

| | | | | |
|---|---|---|---|---|
| | | | | |
| | + | + | | |
| + | - | - | + | |
| | + | + | + | |
| | | | | |

图 7-21　向右旋转 90°后的图像

本文在对轮廓信息进行描述时，利用了轮廓的走势特征设计特征模型。如图 7-19 所示的轮廓，依然选择最上方左边的“+”作为起始点，逆时针表示轮廓，第一、二、三点组成的折线在第二点向左转，接着依次是左转、左转、左转、直行、左转、左转。同时定义了转向的程度，直行用“0”表示，左转小于 90° 表示为“1”，左转 90° 表示为“2”，左转大于 90° 表示为“3”，右转小于 90° 表示为“–1”，右转 90° 表示为“–2”，右转大于 90° 表示为“–3”。对于图 7-19 所示的轮廓图，轮廓表示为：1121021；其向右旋转 90° 后（如图 7-21 所示），轮廓表示为：1121021。两个表示相同，表明本文提出的这种表示方法对字符的旋转具有较好的适应性。

字符图像边缘轮廓走势特征表述为：使用轮廓走势转向及转向程度表示字符边缘轮廓的特征。这种特征在计算时使用图像处理中的轮廓跟踪算法作为第一步计算轮廓；然后利用向量叉乘和点积的正负号判断转向及转向程度。由于在图像处理中，$y$ 轴的正方向一般定义为垂直向下，因此叉乘为负表示左转，为正表示右转。同时叉乘为 0 且点积为正，则表示方向为直行；若叉乘为 0 而点积为负，则表明向后 180°，归为左转大于 90° 一类。特征计算的算法流程如下。

（1）利用轮廓像素跟踪算法提取物体图像的轮廓像素点顺序表 $P_i$，$i$ 的取值范围为 1～$n$。

（2）遍历图像中所有的像素点 $i$，并选择相邻顶点两个像素点组成三元组 $e=(P_{i-1},P_i,P_{i+1})$，由 $P_{i-1}$ 和 $P_i$ 构成向量 $\boldsymbol{a}=(x_1,y_1)$，$P_i$、$P_{i+1}$ 组成向量 $\boldsymbol{b}=(x_2,y_2)$，利用向量 $\boldsymbol{a}$、$\boldsymbol{b}$ 确定三元组 $e$ 在 $P_i$ 点处的转向及转向程度，并赋值为 $V_i$。计算过程如下。

叉积：$C_i=\boldsymbol{a}\times\boldsymbol{b}=x_1\times y_2-x_2\times y_1$

点积：$D_i=\boldsymbol{a}\cdot\boldsymbol{b}=x_1\times y_1+x_2\times y_2$

$V_i$=0，如果 $C_i$=0，且 $D_i$>0；

$V_i$ =1，如果 $C_i$<0，且 $D_i$>0；

$V_i$ =2，如果 $C_i$<0，且 $D_i$=0；

$V_i$ =3，如果 $C_i$≤0，且 $D_i$<0；

$V_i$ =–1，如果 $C_i$>0，且 $D_i$>0；

$V_i$ =–2，如果 $C_i$>0，且 $D_i$=0；

$V_i$ =–3，如果 $C_i$>0，且 $D_i$<0。

如图 7-22 所示，$p_1$ 的坐标为(3,2)，$p_2$ 的坐标为(2,3)，$p_3$ 的坐标为(2,4)。

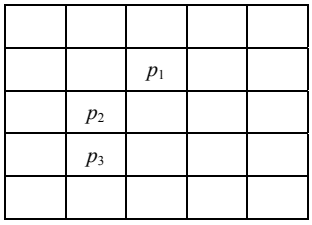



图 7-22　特征计算实例

向量 $\boldsymbol{a}$=$p_1p_2$=(–1,1)，$\boldsymbol{b}$=$p_2p_3$=(0,1)；叉积 $C$=–1×1–1×0=–1；点积 $D$=1，则 $V$ 取值为 1。

轮廓跟踪算法的基本思想如下：

（1）初始化位置点 $p$ 为左上方的目标像素点；初始化轮廓集合 $S$ 包含 $p$ 点，$S$←$p$。

（2）初始方向为左上方 45°。

（3）逆时针扫描 $p$ 点相邻的点，若无目标像素点，则结束；否则执行步骤（4）。

（4）遇到目标字符像素点 $m$，则加入集合 $S$，$S \leftarrow m$，更新 $p$ 的位置为 $m$ 点，执行步骤（2）。

字符边缘轮廓走势特征提取的是字符形状的边缘特征信息，对于字符的完整性要求较高。该算法适用于图片经过预处理后字符损失少且可分割的验证码情况，处理的是单个字符图片。它较为有效地描述了字符的边缘轮廓信息，且能适应字符的旋转。其缺点是对字符完整性要求较高，且对于存在字符拉伸等形变的情况适应性不好。

### 3．像素位图特征

像素位图特征是直接使用字符图像的像素信息，它是一种很简单的、原始的特征信息。其处理方式是：将字符图像归一化处理为大小固定的图片，然后按行扫描，将像素点信息组织为多维向量特征。对于一个像素点，若其像素值为 0，则设置向量对应位置为 0；若像素值为 255，则设置向量对应位置为 1。也就是说，向量特征是由 0、1 值组成的多维向量。如果图片的大小固定为 15×20 像素，则向量的维度是 300。因为向量某一位置的取值空间为 0、1，存储时可以采用位图存储，即用 1bit 存储向量的一个位点，这种方式大大缩减了存储的空间代价。最终在计算特征向量差异时使用异或运算得到存在差异的位点，进而评估差异值。使用位运算极大地提高了运算速度。

像素位图特征使用的是字符图像的全部像素信息，因而描述字符的信息非常全面，对于在预处理时有像素损失的字符，依旧能较好地描述字符，并且能保证不同字符之间的差异性。但是，该特征不能适应字符的旋转和形变，同一字符旋转或变形后的图片像素点位置不再一样，特征差异会较大。同时，它存储的信息量较大，因而空间复杂度及计算时的时间复杂度均相对较大。虽然像素位图特征有上述缺点，但在字符干扰信息难以去除或者去除后像素损失较多、字符不可分割等情况下，该特征在模板匹配中能取得较好的应用效果。

## 7.5.2 特征库生成

特征库生成过程即制作特征模板库，用于在模板匹配过程中与字符特征匹配计算，完成识别过程。特征库包含的特征越多，对字符信息的描述越完备，匹配度就越高。特征库的一条数据包括字符的特征和特征的标签，即字符的文本表示（一个字符），可看作特征的分类。特征库的制作主要包含两个步骤。第一步，根据训练图片样本，提取字符的像素特征信息。提取字符像素特征的过程需要完成字符识别全过程中除了匹配识别步骤外的所有步骤。第二步，为每个特征添加标签。此步骤可以手动完成，也可以先聚类，待提取类别中心特征后再给每个类别添加标签。本文采用了聚类算法思想辅助添加标签。聚类算法思想将特征距离相近的特征分为一类，主要技术有特征向量距离度量和聚类方法设计。

### 1．特征向量距离度量

特征向量之间的距离即具体表示字符之间差异性的量化方法。向量距离常用的有欧式距离（Euclidean Distance）、曼哈顿距离（Manhattan Distance）和向量夹角余弦距离。本文在处理字符区域像素统计特征和字符边缘轮廓走势特征时使用了向量夹角余弦距离。向量夹角余弦距离的公式如下：

$$\mathrm{Dist}=\cos(\vec{\boldsymbol{a}},\vec{\boldsymbol{c}})=\frac{\vec{\boldsymbol{a}}\cdot\vec{\boldsymbol{c}}}{|\vec{\boldsymbol{a}}|\times|\vec{\boldsymbol{c}}|} \tag{1}$$

其中，$\vec{\boldsymbol{a}}=(x_1,x_2,\cdots,x_n)$、$\vec{\boldsymbol{c}}=(y_1,y_2,\cdots,y_n)$，表示两个特征向量。则

$$\mathrm{Dist}=\frac{\vec{\boldsymbol{a}}\cdot\vec{\boldsymbol{c}}}{|\vec{\boldsymbol{a}}|\times|\vec{\boldsymbol{c}}|}=\left(\sum_{i=1}^{n}x_i\cdot y_i\right)\Big/\left(\sqrt[2]{\sum_{i=1}^{n}x_i^2}\sqrt[2]{\sum_{i=1}^{n}y_i^2}\right) \tag{2}$$

本文在处理像素位图特征时，图片大小固定，像素点总数固定，使用像素点差异数量占总像素数的比率作为衡量特征的距离。存储时使用的是按比特位存储，计算距离时第一步使用异或运算求解像素点不一致的位置，第二步统计特征中不一致的像素点的个数并计算比率值。具体的计算步骤如下：

（1）计算像素颜色值不一致的像素点位置信息，使用异或运算。位图特征向量 $\boldsymbol{a}=(x_1,x_2,\cdots,x_n)$，$\boldsymbol{b}=(y_1,y_2,\cdots,y_n)$，其中 $x_i$、$y_i$ 取值为 0 或者 1。1 表示该位置是字符像素点，即值为 255；0 表示该位置是背景点，即值为 0。通过异或运算 $\boldsymbol{a}\text{^}\boldsymbol{b}$，可以将所有位置不相同处设置为 1。

（2）统计异或运算结果中值为 1 的个数，可以使用移位运算和与运算实现。然后计算不同像素点个数所占比率值，公式为：Rate=Count/Sum。其中，Count 为 1 的个数，Sum 为像素点总数。

在特征库制作和识别过程中的匹配环节都用到了特征向量距离度量，它是验证码识别中不可或缺的一种处理技术。

### 2．聚类及设置标签

在制作验证码字符的特征模板时，选择的验证码图片会非常多，在实验中一般一种验证码会选取 2000 张左右的图片，提取出的有效字符图片会达到 8000 张左右，通过对这些字符图片提取特征，然后手动标记，工作量会很大。同时，本文研究的验证码是字母和数字的，字符的个数有限，且较少，特征模型也需要具有泛化表示的能力。因此，将特征进行聚类运算，分成多个类别，然后对类别统一标注，这既能减少手动标记的工作量，也使得特征库符合实现的需要，能适应字符特征的较小变化。

K-Means 算法是一种简单聚类算法，且能取得较好的聚类效果，本文采用它实现聚类过程，主要流程如下：

（1）选择 $k$ 值，即所分类别的数量，依据经验选取。

（2）运行 K-Means 算法将样本字符的特征聚类为 $k$ 类，并提取每个类别中距离最大的两个特征及其对应的字符。

（3）根据选取的 $k$ 对距离最大的字符，若 $k$ 对字符全部相同，则适当减小 $k$ 值；若有较多字符不同，则说明分类数太少，需要适当增加 $k$ 值。确定 $k$ 值后重

新返回步骤（3），运行聚类算法的过程。过程的结束需要依赖前后两次的效果，若前一次字符相同的对数比率极大，而后一次较少到适当值，或者前一次字符相同的对数比率偏小，而后一次得到了较好的提升，则可以结束聚类过程的运行。这一步需要凭经验干预过程执行，目的在于选择较为合适的 $k$ 分类数。

（4）聚类算法结束后，将使用 $k$ 个类别的质心作为类的代表，并根据字符人为设置标签，存入特征模板库中。

使用聚类方法制作模板能适应字符特征较小程度的变化，使用类别中心减少了特征模板的数量。

### 7.5.3 识别方法

验证码的识别步骤是指将待测试的图片数据集或待识别的在线图片的特征与特征库匹配识别出文本字符的过程。其主要操作是匹配操作，以及根据策略选择具有相似性特征对应的标签作为识别结果。本文使用 k 最近邻算法（KNN 算法）为待处理的特征分类，设置标签。KNN 算法的思想是：通过将待分类的特征与特征库里的特征匹配比较，选择最相似的 $k$ 个特征及对应的标签，然后统计各类标签在 $k$ 个标签中的数量，将出现次数最多的标签作为待分类特征的标签值。此步使用了向量间的距离度量方式来计算特征间的相似度。KNN 算法是一种简单的监督型机器学习算法，其设计简单，能满足本文识别过程的需求。

## 7.6 实验结果及分析

我们研究了三种不同的验证码，并分别根据前几节介绍和提出的方法设计了相应的识别过程。基本的验证码图片的识别流程如图 7-23 所示，包括验证码图片获取、图片预处理、字符分割、特征库制作与字符识别 5 个步骤。

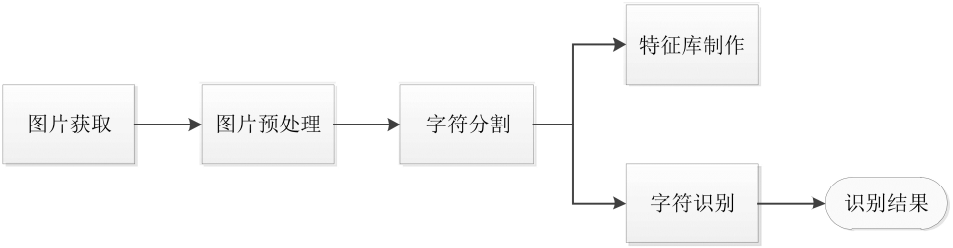



图 7-23　验证码识别流程

我们选择了三种具有代表性的验证码进行实验，在识别中对于每个阶段都能对应使用不同的方法处理，并分析识别的结果等信息。

## 7.6.1　使用轮廓走势特征的识别

### 1．验证码特点分析

本节研究图片色彩多样的验证码识别，如图 7-24 所示，其特点如下：

（1）图片中的字符均为数字，数字不在同一水平线上，在垂直方向有错位。

（2）数字有旋转和形变。旋转程度较小，形变程度较低。

（3）图片有干扰线，干扰线为直线型，且颜色亮度较浅。

（4）字符有颜色干扰，且干扰线会跨越多个字符，导致字符粘连。

（5）数字与数字之间有一定间距，字符之间未连接在一起。

（6）数字存在笔画干扰，如数字 6 中的圈被涂成一块、数字 1 添加边缘线条等。

（7）图片无背景色，较为整洁。

图 7-24　验证码图片

该验证码字符之间不连接，字符边缘保存完整，如果预处理效果较好，则字符可分割开，并且字符无损失，适合使用本文提出的边缘轮廓走势特征表示。

### 2．识别详细流程设计

根据图片的特点，在设计预处理时，图片无背景、无噪声，因此不需要图像去噪和背景去除过程。图片预处理阶段的任务是完成图片干扰线的去除，以及图片的灰度化和二值化。

根据干扰线为直线且亮度较低这一特点，同时分离出各通道比对，发现在 Alpha 通道，数字和干扰线的亮度有较为明显的差异，如图 7-25 所示。因此，可以使用提取 Alpha 通道完成图像灰度化处理和直方图阈值处理方法对图片进行二值化处理。

（a）原图　　（b）Alpha 通道图

图 7-25　图片通道分析

在字符分割阶段，因干扰线去除效果好，预处理输出的结果只包含了数字，数字粘连是因为干扰导致的，数字之间不存在直接粘连，如图 7-26 所示。因此，使用 7.4.4 节设计的连通区检测算法分割字符效果较好。

图 7-26　预处理结果图

字符分割后，字符信息保存完整，因此，在特征建模时采用字符边缘轮廓走势特征模型。本节验证码图片的识别程序主要包括：图片 Alpha 通道提取、直方图阈值、连通区检测字符分割、轮廓走势特征提取、特征库制作、匹配识别。具体流程如图 7-27 所示。

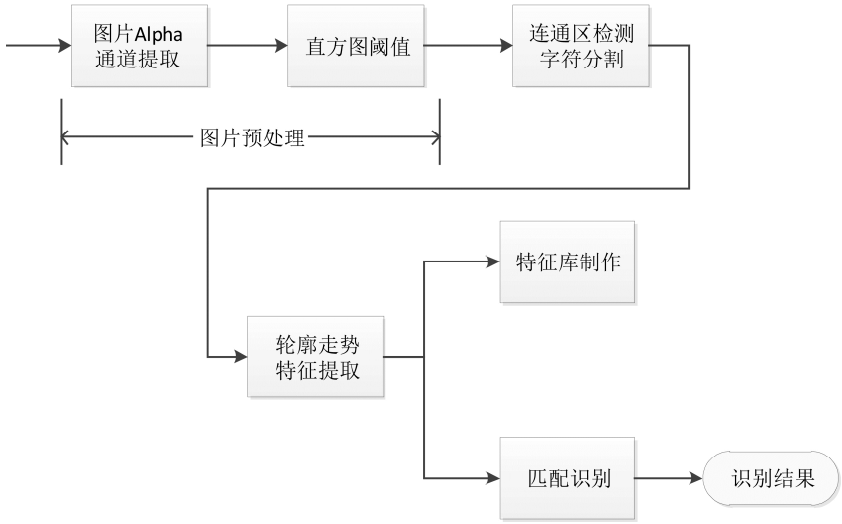


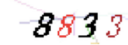
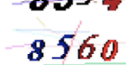
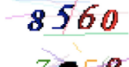

图 7-27　识别流程图

### 3．实验结果

实验中，采用 2000 张验证码图片作为实验数据，其中 1800 张图片用于训练制作特征库，200 张图片用于测试分析识别结果。

部分识别结果如表 7-1 所示。识别中取得了较好的效果，对于字符 2、3、5、6、4、9 基本能够识别正确；在识别 1 和 7 时，可能会因为 7 的形变导致识别为 1，如验证码“6768”；在识别 8 时，可能会因为 8 字符加粗后识别为 0，如验证码“1418”。

表 7-1　部分识别结果

| 验证码 | 识别结果 | 正确结果 | 验证码 | 识别结果 | 正确结果 |
|---|---|---|---|---|---|
|  | 6470 | 6470 |  | 2263 | 2263 |
|  | 6168 | 6768 |  | 8833 | 8833 |
|  | 8554 | 8554 |  | 7155 | 1155 |
|  | 8560 | 8560 |  | 1410 | 1418 |
|  | 7658 | 7658 |  | 2607 | 2607 |

使用上文提及的识别方法，将 1800 张图片使用轮廓走势特征制作特征库，且依次使用的图片数量为 600、800、1000、1200、1400、1600、1800，共 7 种方式，并依次使用 7 种特征库对 200 张图片进行识别处理。识别的准确率计算公式如下：

$$\text{rate} = \frac{m}{N} \tag{3}$$

实验中计算了两种准确率。一是字符识别的准确率，每张图片有 4 个字符，此时，$m$ 代表识别正确的字符个数，$N$ 代表图片中字符的总个数，$N$=4$P$，$P$ 为图片的张数，$P$=200。二是验证码图片的识别准确率，即识别正确的图片张数所占的比率，只有当图片的 4 个字符均识别正确时，验证码图片识别才正确，此时，$m$ 代表识别正确的图片张数，$N$ 为图片总数，$N$=200。识别结果如图 7-28 所示。

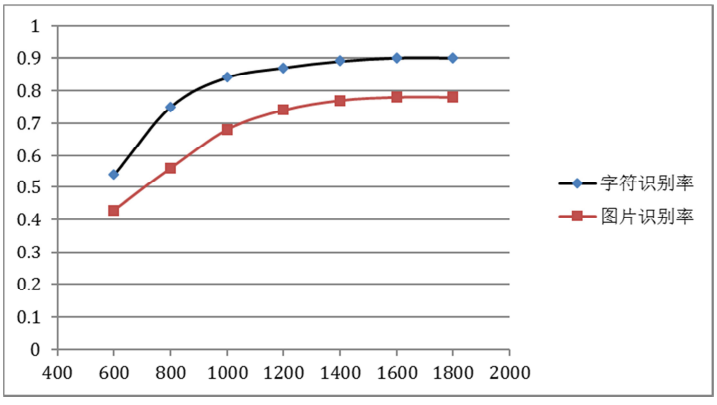



图 7-28　识别率与图片张数的关系

由图 7-28 可知，对于本节研究的验证码，本文设计的识别方法取得了较好的识别效果，字符识别率为 90%，验证码图片识别率为 78%（训练图片张数为 1800 张时）。同时，在训练图片数量为 1200 张之前，识别率随着训练图片数量的增加不断增加，在 1200 张之后，识别率趋于稳定，说明针对本节研究的验证码，1200 张图片能够提供较为全面的特征库。这表明针对该验证码，本文设计的字符轮廓

走势特征能较好地对字符建模。

如图 7-29 所示为对本节验证码分别采用区域像素统计特征和轮廓走势特征的识别率。可以看到，识别率均随着制作特征库图片的数量增加而增长，然后趋于稳定。区域像素统计特征的识别率在 1400 张图片时趋于稳定，轮廓走势特征在 1200 张左右区域稳定，且在稳定前，轮廓走势特征的识别率均高于区域像素统计特征，这表明轮廓走势特征能在特征图片库较小时取得较高的识别率，更能适应字符的旋转和形变。另外，最终两种特征的稳定识别率相近，且识别率均为 78%，均能对本节研究的验证码有较好的识别效果。在两种特征均能识别时，本文提出的轮廓走势特征效果明显。

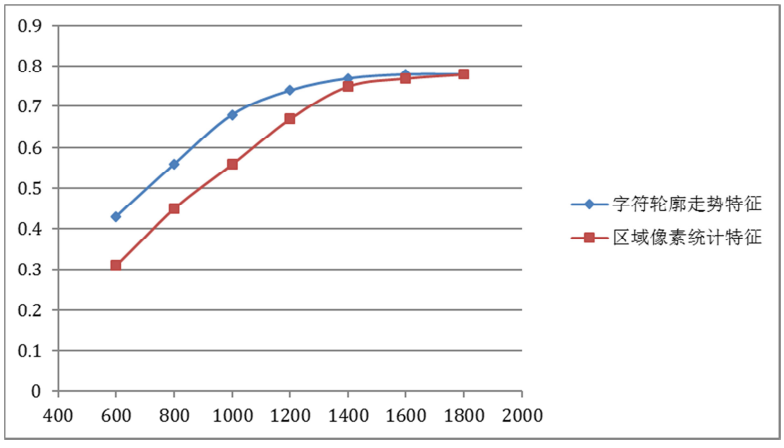



图 7-29　区域像素统计特征和轮廓走势特征识别率

## 7.6.2　分割并使用统计特征的识别

### 1. 验证码特点分析

本节研究图片色彩单一、带曲线干扰线的验证码识别，如图 7-30 所示。其特点如下：

（1）图片色彩单一，字符与干扰线使用相同的颜色和亮度，背景使用另一种颜色，颜色直方图存在双峰性质，可以使用 OTSU 二值化技术。

（2）字符存在旋转变换、缩放变换，但没有形状变换，字符特征种类较少。

（3）干扰线与字符颜色和亮度相同，且是曲线，字符消去干扰线后像素会存在损失。

（4）干扰线造成多个字符粘连，同时字符之间存在轻微的直接连接，字符分割难度较大。

（5）字符不在同一水平线上，且字符大小不一致。

图 7-30　验证码图片示例

字符去干扰线会有像素损失，且字符分割的难度大，分割后的字符存在不完整的情况，不适合使用轮廓走势特征，在特征建模时本文使用区域像素统计特征模型。

### 2．识别详细流程设计

根据图片背景与字符颜色和亮度单一的特点，同时图片无噪声点，因此，在字符预处理阶段只需包含验证码图片灰度化处理、二值化分类和曲线干扰线去除等步骤。

图片颜色较为单一，在灰度化处理时将 RGB 颜色空间变换到 YUV 颜色空间，并以亮度（Y）分量值作为灰度值。灰度化处理后，图片的灰度直方图具有双峰性质，使用最大类间方差的阈值方法进行二值化处理。

该类验证码的干扰线为曲线，且和字符有连接，针对本节的验证码干扰，本文根据干扰线的特征设计曲线干扰线去除方法。图片预处理的结果如图 7-31 所示。

（a）原图　　（b）预处理结果图

图 7-31　验证码图片预处理

在字符分割阶段，因预处理会造成字符像素点损失，字符可能被分开。同时字符之间存在粘连，造成字符分割的难度较大。我们针对该验证码的特点，设计了基于连通区域和投影的字符分割算法，它结合了本文改进的连通区检测和投影分割的优点，并增加了分割后处理的方式，适应本文验证码的字符分割。字符分割的结果如图 7-32 所示，分割效果较好。

（a）原图　　（b）分割结果

图 7-32　字符分割

在字符特征建模阶段，因为字符的分割结果有像素的损失，所以本文使用字符区域像素统计特征，该特征计算的区域的像素比率不需要字符图片大小固定。

本节的验证码识别流程包括：基于颜色空间变换的灰度化、使用最大类间隔阈值进行二值化处理、曲线干扰线去除、基于改进的连通区检测和投影的字符分割、区域像素统计特征建模、特征库制作、字符匹配识别。具体流程如图 7-33 所示。

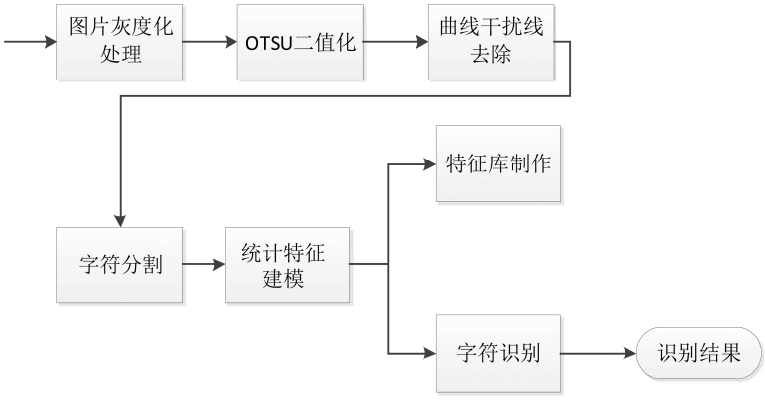



图 7-33　识别流程图

### 3．实验结果

采用 2000 张验证码图片作为实验数据，其中 1800 张作为实验训练数据，制作特征模板库，剩余 200 张作为测试数据。本节研究的验证码的识别难点在于字符分割，使用基于连通区检测和投影的分割方法，并利用区域像素统计特征建模。

部分识别结果如表 7-2 所示。可以看到，分割结果能基本将字符分割开，干扰线去除导致部分像素损失，会使得识别结果变差，比如验证码“mhku”，k 的损失较重，致使识别为 1。本文设计的字符分割算法对于字符存在像素损失的情况具有较好的适应性，如验证码“syt2”因去除干扰线导致 y 不连续、验证码“mhku”中字符 h 不连续等，在分割时都取得了较好的分割效果。

表 7-2　部分识别结果

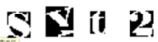
| 验证码 | 分割结果 | 识别结果 | 正确结果 |
|---|---|---|---|
|  |  | uff7 | uff7 |
|  |  | cy7h | cy7h |
|  |  | syt2 | syt2 |
|  |  | mh1u | mhku |

图 7-34 展示了本节实验识别率与制作特征库图片数量的关系。可以看出，本文设计的识别流程和提出的分割方法在本节的验证码识别中取得了显著的效果，字符识别率达到 63%，验证码图片识别率达到 51%（特征库图片数量为 1800 张时）。

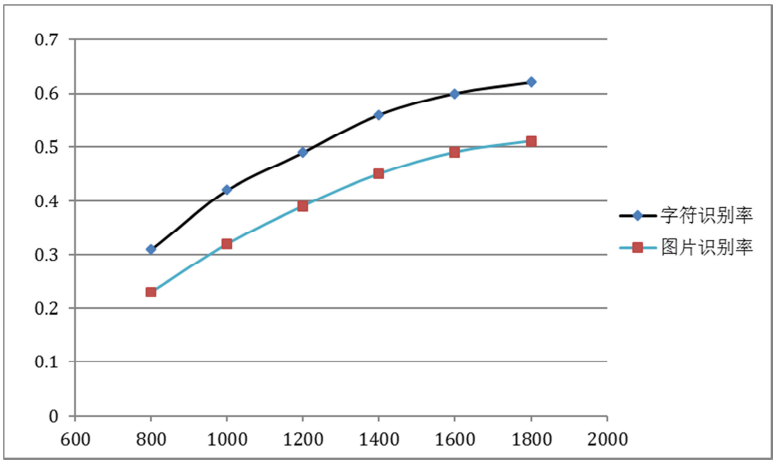



图 7-34　识别率与特征库图片数量的关系

另外，由图 7-34 可知，识别率均随着训练集中图片数量的增多而增大，且在 1400 张之前识别率增加的幅度较大，在 1400 张之后增长较慢。这说明当图片数量达到 1400 张时，利用图片制作的特征基本覆盖了绝大部分字符图片；在 1400 张之后识别率依然增大，说明依旧有新的特征模板加入，干扰线和字符粘连影响了特征的提取。

## 7.6.3　不分割且使用位图特征的识别

### 1. 验证码特点分析

此处研究了字符之间粘连程度较大的验证码识别，如图 7-35 所示。其特点如下：

（1）无背景颜色，无噪声，无干扰线。

（2）字符间存在较大程度的粘连。

（3）字符有旋转和形变，且形变程度较大。

（4）图片中存在单个字符，可以用作特征模板制作。

（5）字符个数不固定，有 3 个和 2 个两种情况。

图 7-35　验证码图片

该验证码在字符分割上难度非常大，原因是字符旋转与字符形变导致字符粘连的形式多样、字符粘连程度较大等造成设计分割方法困难。我们避开设计字符分割，直接按模块宽度提取像素，建模匹配的方式，设计识别方法。建模方式采用像素位图特征。

### 2．识别详细流程设计

此处研究的验证码图片无背景颜色、无噪声、无干扰线。因此，图片预处理阶段仅包括图片颜色像素灰度化处理和图像二值化处理两个步骤。

图片灰度化采用在 YUV 空间中提取亮度分量作为灰度值的方法。图片二值化处理时采用普通阈值处理，阈值为 200。图片预处理结果如图 7-36 所示，效果较好。

（a）原图　　（b）预处理结果

图 7-36　图片预处理

对于预处理后的图片，使用改进的连通区检测算法进行初步分割，图片中存在单个字符，在制作模板库时，使用分割的单个字符建立模板。在识别过程中，对于单个字符直接使用匹配的方式识别；对于存在粘连的多个字符（2 个或 3 个），从左右两边分别与模板匹配，选择匹配度最高的识别方式。匹配中使用的特征是像素位图特征。匹配识别的具体流程如下：

（1）对于分割后的连通区域，如果宽度符合单个字符的宽度范围，则在计算位图特征后直接与模板库中的特征匹配，利用 KNN 算法的思想确定识别结果。

（2）如果连通区宽度较大，左边的区域与模板库匹配，选择与模板同宽度的区域计算位图特征，匹配计算相似度，选择相似度最大的匹配方式。然后右边的区域与模板库匹配，识别字符。根据验证知识，验证码中的字符最多为 3 个，可判断区域字符的个数，若为 3 个，则需要将中间剩余的像素与模板匹配识别。

本节验证码识别的主要流程包括：图片灰度化处理、二值化处理、连通区检测字符分割、特征库制作、单字符匹配识别、多字符识别。具体流程如图 7-37 所示。

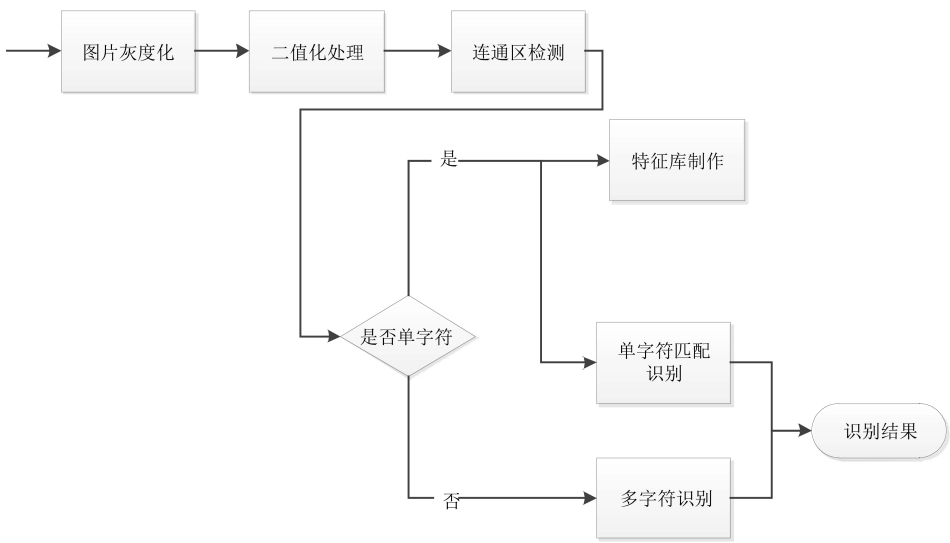



图 7-37　识别流程图

### 3．实验结果

本实验主要研究粘连且字符有形变和旋转的验证码，且粘连程度较大，研究了本文设计的识别方法对于粘连程度较大的验证码识别的可行性。本实验中，在字符分割阶段使用了我们设计的改进的连通区检测算法尽量分割，对于已分割开的单字符，在训练阶段用于建立特征库，在识别阶段直接用于匹配识别；对于不能分割的多字符连通区域，采用上一小节设计的策略进行匹配识别。

实验数据为从网站下载的 2000 张验证码图片，其中 1800 张用于训练时制作特征库，200 张作为测试图片集。部分识别结果如表 7-3 所示。可以看到，本文设计的方法对验证码有较为准确的识别结果，对于粘连较大且形变较为严重的验证码识别率较低。

表 7-3　部分识别结果

| 验证码 | 分割结果 | 识别结果 | 正确结果 |
|---|---|---|---|
|  |  | 75 | 75 |
|  |  | 732 | 732 |
|  |  | 899 | 899 |
|  |  | 33 | 33 |
|  |  | 733 | 734 |

图 7-38 为识别率与训练时（制作特征库）图片数量的关系图。可以看到，本节设计的识别方法对于粘连且有旋转和形变的验证码具有较好的识别能力，字符识别率达 42%，验证码图片的识别率达 22%（图片数量为 1800 张时）。

同时，由图 7-38 可知，字符识别率和图片识别率均随着图片数量增加而提高，但是图片识别率增加的幅度远小于字符识别率增加的幅度，这说明字符的粘连程度大大提高了一张图片中存在字符识别错误的概率，表明字符粘连能提高验证码的反自动识别能力。

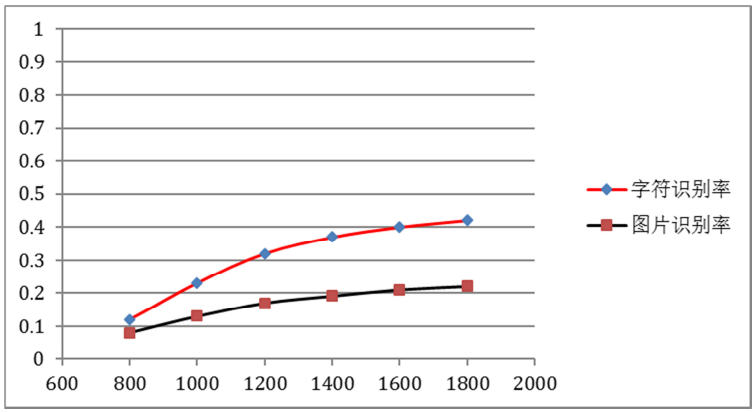



图 7-38　识别率与训练数据图片数量的关系

## 7.7　验证码识别理论和技术在国内外的研究现状

验证码识别在国内外均有许多学者进行研究，取得了大量卓有成效的研究成果。验证码识别和字符识别具有较大的相似性，识别过程中可以借鉴参考已有的研究成果。但是，验证码识别有其独特性。验证码设计的本身就是反识别的，安全性一直是设计验证码追求的目标之一，所以验证码识别比字符识别难度更大。验证码图片通常都加入了干扰因素，如背景色、噪声点、干扰线等，字符选择、字符粘连、字符形变等都增加了识别的难度。

验证码有其优势与缺陷，设计者正是利用验证码的优势防止自动代码或程序的识别，保证其安全性。2011 年，Bursztein、Elie 等[39]人分析了现有验证码的优势和缺点，提出了设计验证码的一些规则，包括反分割的方法，如粘连、旋转等，这些规则也是现今验证码设计者常用的反识别技术。Edward Aboufadel、Julia Olsen 和 Jesse Windle[8]通过把待识别的字符旋转操作至水平方向，进而使用小波滤波结果为特征，破解了 the Holiday InnPriority Club 的验证码。小波滤波是图像处理在图像频域变换的常用方式，这里利用了验证码图片在频域空间的特征。机器学习的方法也在验证码识别中取得了很多成果，Mittal A 和 Kumar A[40]使用了 SVM 识

别验证码。还有许多学者对验证码图片的特征进行了深入分析，提出了效果明显的识别方法，如 Belongie S 和 Malik J[10]使用了形状上下文的方法识别验证码，对于字符的旋转和形变有较强的抗性。验证码识别可分为多个阶段，多数研究的重点主要集中在图片字符的分割和字符识别阶段。Kun Fang、Zhang Bu 和 Xia Zheng you[11]利用复杂网络设计了一种改进的社区划分模型，对验证码的图片文字先分割再识别。2007 年，Jeff Yan 和 Ahmad Salah El Ahmad[12]设计了朴素模式识别算法破解 Captchaservice.org 提供的验证码。2012 年，Korayem M、Mohamed A A 等[13]利用计算机视觉的特征，包括汇总统计特征、灰度直方图特征、原始像素值的 50×75 维向量特征、方向梯度直方图特征、GIST 描述子，设计了学习分类算法，在 Avatar Captcha 验证码系统上获得了效果明显的识别能力。Cruz-Perez C, Starostenko O 等[14]利用字符的形态学特征，设计了启发式的字符分割算法识别 reCAPTCHA 的验证码，分割的成功率达到 82%，识别的准确率达到 94%。字符分割是验证码识别过程中的重要步骤，Tingre S 和 Mukhopadhyay D[15]研究了字符的分割算法，并实践了垂直分割算法，在 ez-gimpy 验证码上的分割准确率为 100%，IDEA 和 MSRTC 分别为 75%和 40%。

国内也有许多学者对验证码进行了研究。贺强和晏立[[16]基于字符形状上下文的特征提出了改进的方法用于识别复杂的验证码系统，这种方法避免了分割对字符的损害，对粘连较为严重的验证码有较好的识别率。王璐、张荣等[17]使用卷积神经网络的方法进行粘连字符的图片识别，并在西祠胡同和猫扑的验证码系统上测试，获得了较高的识别率。王虎、冯林等[18]使用模板匹配的算法，分析了验证码的识别方法，并且提出和分析了一种加权模板的构造方法。尹龙、尹东等[19]对粘连字符采用了基于密集尺度不变特征变换和随机抽样一致性算法识别的方法。2011 年，Huang Shuguang、Zhang Liang 等[20]使用递归神经网络，基于全面验证的方法设计了一种识别算法，该算法能够识别粘连的验证码，同时提高了递归神经网络的识别率。

## 7.8　本章小结

验证码是互联网安全的基本手段之一，其广泛应用于网站的注册、登录等基础服务的安全防护上。本章针对现在应用广泛的文字图片类验证码进行了分析与研究，选取了具有代表性的三种验证码，分析其特点和缺点，并具有针对性地设计和实现了识别的流程，取得了较好的识别效果。

但是，验证码的技术形式多样，本文研究的干扰信息和字符并行处理的方式较为常见，且并非难度极大的方式，在图片预处理的技术手段上需要进一步研究，对于噪声和干扰线等研究更为合理的技术，尽量保全字符的像素信息。

在识别过程中，本章采用的是字符识别最常用的模板匹配方式，如何使用现在应用广泛的机器学习技术，如神经网络和深度学习技术等是当前的研究热点。同时，在对图片字符建模时，如何能更加客观地描述字符，适应字符的旋转和形变，是字符识别中需要合理研究与设计的。

## 参考文献

[1]　侯玉锋．粘连字符验证码识别关键技术研究[D]．北京邮电大学，2016．

[2]　Serge Belongie, Greg Mori, Jitendra Malik. Matching with Shape Contexts[M]// Statistics and Analysis of Shapes. Birkhäuser Boston, 2006:81-105.

[3]　张云刚，张长水．利用 Hough 变换和先验知识的车牌字符分割算法[J]．计算机学报，2004，27（1）：130-135．

[4]　陈寅鹏，丁晓青．复杂车辆图像中的车牌定位与字符分割方法[J]．红外与激光工程，2004，33（1）：29-33．

[5] Anil K. Jain. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8):651-666.

[6] Baird H S, Riopka T P. ScatterType: a reading CAPTCHA resistant to segmentation attack[J]. Proc Spie, 2005:16-20.

[7] Bursztein E, Martin M, Mitchell J. Text-based CAPTCHA strengths and weaknesses.[C]. Acm Conference on Computer & Communications Security. 2011:125-138.

[8] Aboufadel E, Olsen J, Windle J. Breaking the Holiday Inn Priority Club: CAPTCHA.[J]. College Mathematics Journal, 2005, 36:101-108.

[9] Mittal, Akshay and Ankit Kumar. Recognising Text from CAPTCHA.

[10] Belongie S, Malik J. Matching with Shape Contexts[D]. Berkeley,USA: University of California at Berkeley, 2000.

[11] Fang K, Bu Z, Xia Z Y. Segmentation of CAPTCHAs based on complex networks[J]. Lecture Notes in Computer Science, 2012, 7530:735-743.

[12] Yan J, El Ahmad A S. Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms[C]. Computer Security Applications Conference, 2007. ACSAC 2007. Twenty-Third Annual. IEEE, 2007:279-291.

[13] Korayem M, Mohamed A A, Crandall D, et al. Learning visual features for the avatar captcha recognition challenge[C]. Machine Learning and Applications (ICMLA), 2012 11th International Conference on. IEEE, 2012, 2: 584-587.

[14] Cruz-Perez C, Starostenko O, Uceda-Ponga F, et al. Breaking reCAPTCHAs with Unpredictable Collapse: Heuristic Character Segmentation and Recognition[J]. Lecture Notes in Computer Science, 2012, 7329:155-165.

[15] Tingre S, Mukhopadhyay D. An approach for segmentation of characters in CAPTCHA[C]. Computing, Communication & Automation (ICCCA), 2015 International Conference on. IEEE, 2015.

[16] 贺强，晏立．基于形状上下文的复杂验证码识别算法[J]．计算机工程，2011，37（2）：200-202．

[17] 王璐，张荣，尹东，等．粘连字符的图片验证码识别[J]．计算机工程与应用，2011（28）：150-153．

[18] 王虎，冯林，孙宇哲．数字验证码识别算法的研究和设计[J]．计算机工程与应用，2007（32）：86-88．

[19] 尹龙，尹东，张荣，等．一种扭曲粘连字符验证码识别方法[J]．模式识别与人工智能，2014（3）：235-241．

[20] Shuguang H, Liang Z, Pengpo W, et al. A CAPTCHA Recognition Algorithm Based on Holistic Verification[C]. 2011 International Conference on Instrumentation, Measurement, Computer, Communication and Control. IEEE Computer Society, 2011:525-528.

[21] 冈萨雷斯[美]．数字图像处理（MATLAB 版）[M]．北京：电子工业出版社，2006．

[22] 李建华，马小妹，郭成安．基于方向图的动态阈值指纹图像二值化方法[J]．大连理工大学学报，2002，42（5）：626-628．

[23] 孙少林，马志强，汤伟．灰度图像二值化算法研究[J]．价值工程，2010，29（5）：142-143．

[24] 冯超．K-Means 聚类算法的研究[D]．大连理工大学，2007．

[25] 迟晓君，孟庆春．基于投影特征值的车牌字符分割算法[J]．计算机应用研究，2006，23（7）：256-257．

[26] 刘奇琦，龚晓峰．一种二值图像连通区域标记的新方法．计算机工程与应用，2012，48（11）：178-180．

[27] Congedo G, Dimauro G, Impedovo S, et al. Segmentation of numeric strings[C]. 2013 12th International Conference on Document Analysis and Recognition. IEEE Computer Society, 1995:1038-1038.

[28] 常丹华，何耘娴，苗丹．中英混排文档图像粘连字符分割方法的研究[J]．激光与红外，2010，40（12）：1369-1373．

[29] 李兴国，高炜．基于滴水算法的验证码中粘连字符分割方法[J]．计算机工程与应用，2014（1）：163-166．

[30] 马俊莉，莫玉龙，王明祥．一种基于改进模板匹配的车牌字符识别方法[J]．小型微型计算机系统，2003，24（9）：1670-1672．

[31] 王敏，黄心汉，魏武，等．一种模板匹配和神经网络的车牌字符识别方法[J]．华中科技大学学报（自然科学版），2001，29（3）：48-50．

[32] Bhowmik T K, Ghanty P, Roy A, et al. SVM-based hierarchical architectures for handwritten Bangla character recognition[J]. International Journal on Document Analysis & Recognition, 2009, 12(2):97-108.

[33] 申家振，张艳宁，刘涛．基于形状上下文的形状匹配[J]．微电子学与计算机，2005，22（4）：144-146．

[34] 王建平，盛军，朱程辉．基于小波分析的视频图像字符特征提取方法研究[J]．微电子学与计算机，2002，19（5）：51-53．

[35] 何斌．Visual C++数字图像处理[M]．北京：人民邮电出版社，2001．

[36] 李仲荣．区域表示：线性四元树转换成边界链码[J]．计算机学报，1990（7）：498-506．

[37] 彭铁根，吴惕华．基于边界链码的幅度谱图像识别研究[J]．计算机仿真，2004，21（8）：14．

[38] Rivera M, Mayorga P P. Quadratic Markovian Probability Fields for Image Binary Segmentation[C]. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007:1-8.

[39] Bursztein E, Martin M, Mitchell J. Text-based CAPTCHA strengths and weaknesses.[C]. Acm Conference on Computer & Communications Security. 2011:125-138.

[40] Mittal, Akshay and Ankit Kumar. Recognising Text from CAPTCHA.